

A Multimodal Recurrent Neural Networks for Generating Novel Sentence Level Image Descriptions

Anonymous Author(s)

Affiliation

Address

email

Abstract

In this paper, we present a multimodal Recurrent Neural Network (m-RNN) framework for three multimodal tasks: sentence generation, sentence retrieval given query image, and image retrieval given query sentence. Unlike some previous methods that map both the image and sentence features into a same space, we directly model the probability distribution of generating a word given previous words. We can obtain novel image descriptions by sampling from the word probability distribution. The probability of generating a sentence given an image can also be calculated and used for retrieval tasks. The effectiveness of our model is validated on three benchmark datasets (IAPR TC-12 [8], flickr 8K [27], flickr 30K [13]). Our model outperforms the state-of-the-art generative methods by a large margin. In addition, for the retrieval tasks, we show that our model achieves non-trivial performance improvement over previous mapping methods which directly optimize the ranking objective function for retrieval.

1 Introduction

Generating sentence level descriptions for images is becoming an important task recently and has many applications, such as early childhood education, image retrieval, and navigation for the blind. Thanks to the rapid development of the computer vision and natural language processing technologies, there are more and more works appear recently for this task (see a brief review in Section 2). Many of them treat it as a ranking task. They extract features for both sentences and images, and map them to the same semantic embedding space. These methods well handled the tasks of retrieval the sentences given the query image or retrieval the images given the query sentences. But they can only label query images with the sentences annotation of the images in the datasets, thus lack the ability to describe new image that contains previously unseen combinations of objects and scenes.

In this work, we propose a multimodal Recurrent Neural Networks (m-RNN) model to address both the task of generating novel sentences descriptions for images, and the task of image and sentence retrieval. The whole m-RNN architecture contains a language model part, an image part and a multimodal part. The language model part will learn the dense feature embedding for each word in the dictionary and store the semantic context in a recurrent layer. The image part contains a deep Convolutional Neural Network (CNN) [17] to extract image features. The multimodal part connect the language model and the deep CNN together. We adopt a perplexity based cost function (see details in Section 4) for training the network and the errors can be backpropagated to the three parts to updating the parameters. To the best of our knowledge, this is the first work that incorporates the Recurrent Neural Network in a deep multimodal architecture.

In the experiments, we validate our model on three benchmark datasets: IAPR TC-12 [8], flickr 8K [27], and flickr 30K [13]. we show that our method outperforms the state-of-the-art methods in both the task of generating sentences and the task of image and sentence retrieval by a large margin when





				
Retr.	1. Top view of the lights of a city at night, with a well-illuminated square in front of a church in the foreground; 2. People on the stairs in front of an illuminated cathedral with two towers at night;	1. Tourists are sitting at a long table with beer bottles on it in a rather dark restaurant and are raising their beer glasses; 2. Tourists are sitting at a long table with a white table-cloth in a somewhat dark restaurant;	1. A dry landscape with light brown grass and green shrubs and trees in the foreground and large reddish-brown rocks and a blue sky in the background; 2. A few bushes at the bottom and a clear sky in the background;	1. Group picture of nine tourists and one local on a grey rock with a lake in the background; 2. Five people are standing and four are squatting on a brown rock in the foreground;
Gen.	A square with burning street lamps and a street in the foreground;	Tourists are sitting at a long table with a white table cloth and are eating;	A dry landscape with green trees and bushes and light brown grass in the foreground and reddish-brown round rock domes and a blue sky in the background;	A blue sky in the background;

Figure 1: Examples of the retrieved (we showed the top two) and generated sentences given the query image from IAPR TC-12 dataset. The sentences can well describe the content of the images. We show a failure case in the fourth image, where the model mistakenly treat the lake as the sky.

using the same image feature extraction networks. Our model is flexible and has the potential to be further improved when adopting more powerful deep networks for image and the language.

2 Related Work

Deep structure for image and natural language. The deep neural network structure develops rapidly in recent years for both image and natural language. For images, Krizhevsky et. al [17] proposed a 8 layers deep convolutional neural networks (denoted as AlexNet) for image classification tasks that outperforms previous methods by a large margin, and is widely used in computer vision field. Recently, Girshick et. al [7] proposed a detection framework based on AlexNet. For natural language, the Recurrent Neural Network shows the state-of-the-art performance in many tasks, such as speech recognition and word embedding learning [21, 22, 23].

Image-sentence retrieval. Many works treat the task of describe images as a retrieval task and formulate the problem as a ranking or embedding learning problem [12, 6, 29]. They will first extract words and sentences feature embedding (e.g. Socher et.al [29] uses dependency tree Recursive Neural network to extract sentence features) and image features. Then they optimize a ranking cost to learn an embedding model that maps both the language feature and the image feature to a common semantic feature space. In this way, they can directly calculate the distance between image and sentence. Most recently, Karpathy et.al [15] showed that object level image features based on detection results will generate better results than image features at the global level.

Generating novel sentence description for images. There are generally two categories of methods for this task. The first category assume a specific rule of the language and parse the sentence to divide it into several components [24, 10]. They then associate each components to different objects or attributes in the images (e.g. [18] uses a Conditional Random Field model and [5] uses a Markov Random Field model). This kind of method will generate sentences that are syntactically correct. Another category of methods are generative model of the data. They can generate sentences with richer and more flexible structure than the first group. Previous methods will learn a probability density over the space of multimodal inputs (i.e. sentences and images), using Log-BiLinear model [16], Deep Boltzmann Machines [30], and topic models [1, 14]. Our model fails into this category. The probability to generate the sentence given image can serves as the affinity distance for retrieval.

3 Model Architecture

3.1 Basic recurrent neural network

We briefly introduce the basic Recurrent Neural Network (RNN) [4, 21, 22] that is widely used for many natural language processing tasks, such as speech recognition. Its architecture is shown in Figure 2(a). It has three types of layers in each time session: the input word layer w , the context

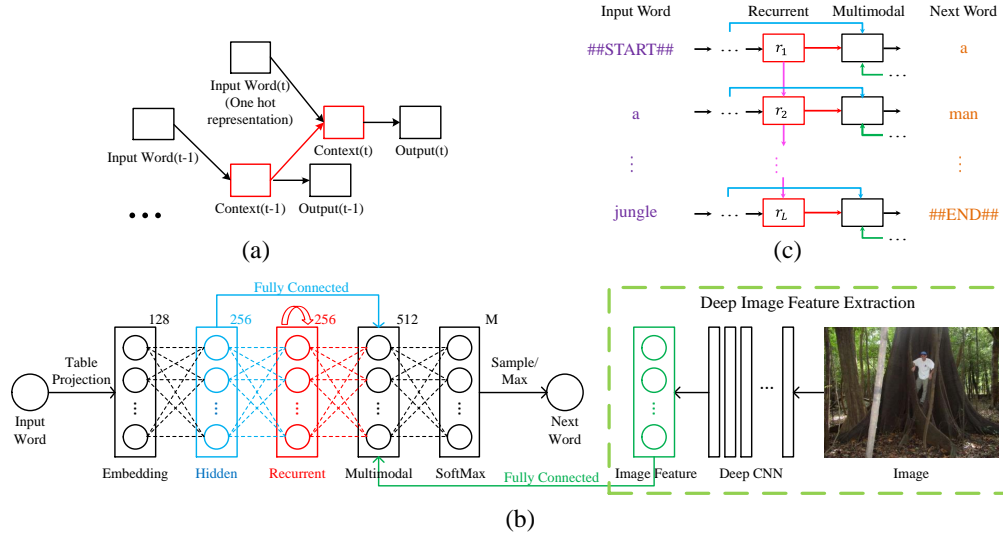


Figure 2: Illustration of the basic Recurrent Neural Network (RNN) and our multimodal Recurrent Neural Network (m-RNN) architecture. (a). The basic RNN architecture. (b). The architecture of m-RNN model. The input of the model is an image and its corresponding sentences. (e.g. the sentence for the shown image is: *a man at a giant tree in the jungle*). The model will estimate the probability distribution of the next word given previous words and the image. This architecture is deeper than the basic RNN. (c). The illustration of how the recurrent layer works in m-RNN. The model parameters are shared for each temporal session of the unfolded m-RNN model.

layer s and the output layer y . The activation of input, context and output layers in time t is denoted as $\mathbf{w}(t)$, $\mathbf{s}(t)$, and $\mathbf{y}(t)$. $\mathbf{w}(t)$ is the one-hot representation of the current word. This representation is binary, and has the same dimension of the vocabulary size with only one non-zero element. $\mathbf{y}(t)$ can be calculated as follows:

$$\mathbf{x}(t) = [\mathbf{w}(t)^T \ \mathbf{s}(t-1)^T]^T; \quad \mathbf{s}(t) = f_1(\mathbf{U} \cdot \mathbf{x}(t)); \quad \mathbf{y}(t) = g_1(\mathbf{V} \cdot \mathbf{s}(t)); \quad (1)$$

where $\mathbf{x}(t)$ as a vector that concatenates $\mathbf{w}(t)$ and $\mathbf{s}(t-1)$, $f_1(\cdot)$ and $g_1(\cdot)$ are element-wised sigmoid and softmax function respectively.

We define two types of “depths” for a RNN architecture. We denote the number of layers in each time session as the *structure depth* (e.g. the structure depth of the basic RNN is 3). When we compare the depth of different RNN networks, we refer to the structure depth. The second depth is denoted as the *temporal depth*. Accordingly, when we do the backpropagation, we need to propagate the error through recurrent connections back in time [28].

3.2 Our m-RNN model

The structure of our multimodal Recurrent Neural Network (m-RNN) is shown in Figure 2(b). The m-RNN model is much deeper than the basic RNN model. Its structure depth is 7 (i.e. input word layer, word embedding layer, hidden layer, recurrent layer, multimodal layer, softmax layer and the next word layer).

The word embedding layer learns a initial dense word embedding representation to replace the one-hot representation in basic RNN. It has several advantages. Firstly, it will significantly lower the number of parameters in the networks because the dense word vector (128 dimension) is much smaller than the one-hot word vector. Secondly, the dense word embedding encodes the semantic meanings of the words. We can find the semantically relevant words by calculating the Euclidean distance between two dense word vectors.

Most of the sentence-image multimodal models [15, 6, 29, 16] use pre-computed word embedding vectors as the initialization of their model. In contrast, we randomly initialize our word embedding layers and learn them from the training data. We show that this random initialization is enough for

our architecture to generate the state-of-the-art results. To further refine the word representation, we add a hidden layer after the initial word embedding layer and treat the activation of the hidden layer as the final word representation.

After the hidden layer, we have a recurrent layer with 256 dimensions. The calculation of the recurrent layer is slightly different from the basic RNN. Instead of concatenating the word representation in time t and the recurrent layer vector in time $t - 1$, we first map the word representation in time t and context in time $t - 1$ into a same vector space and add them together:

$$\mathbf{s}(t) = f_2(\mathbf{U}_w \cdot \mathbf{w}(t) + \mathbf{U}_s \cdot \mathbf{s}(t - 1)); \quad (2)$$

where \mathbf{s} and \mathbf{w} denotes the recurrent layer vector and the word representation respectively.

Inspired by the recent success of the Rectified Linear Unit (ReLU) in training very deep structure in computer vision field [17], we replace the element-wised sigmoid function in basic RNN with ReLU function. ReLU is faster, and harder to saturate or overfit the data than many non-linear functions, such as sigmoid. Previous methods [21, 22] need to conduct backpropagation through time (BPTT) [28] for RNN suffers from the vanishing gradient problem because even the simplest the RNN model has a large temporal depth. They need to use some heuristics, such as truncated BPTT, to avoid this problem. Truncated BPTT will stop the BPTT after k time steps, where k is a hand-defined hyperparameter. Because of the good properties of ReLU, we do not need to stop the BPTT at early stage, which leads to a better and more efficient utilization of the data than truncated BPTT.

After the recurrent layer, we set up a 512 dimensional multimodal layer that connect the language model part and the image part of the whole m-RNN model. The language model part includes the hidden layer (the word representation) and the recurrent layer (the language context). The image part contains the image features. Here we connect the seventh layer of AlexNet [17] to the multimodal layer (please refer to Section 5 for more details). But our framework can utilize any image features. We map the feature vector for each layer to a same feature space and add them together to obtain the feature vector for the multimodal layer:

$$\mathbf{m}(t) = g_2(\mathbf{V}_w \cdot \mathbf{w}(t) + \mathbf{V}_s \cdot \mathbf{s}(t) + \mathbf{V}_I \cdot \mathbf{I}); \quad (3)$$

where \mathbf{w} denotes the multimodal layer feature vector, \mathbf{I} denotes the image feature, $g_2(\cdot)$ is the element-wised scaled hyperbolic tangent function:

$$g_2(x) = 1.7159 \cdot \tanh\left(\frac{2}{3}x\right) \quad (4)$$

This function will force the gradients into the most non-linear value range and can accelerate the training process than the basic hyperbolic tangent function [19].

As the basic RNN, our m-RNN model has a softmax layer that will generate the probability distribution of next word. The dimension of this layer is the vocabulary size M , which is different for different datasets.

4 Training Objective

We use the average logarithm of the *Perplexity* of the sentences in the training set given their corresponding images as the cost function of our m-RNN model. Perplexity is a standard approach to evaluating language model. The perplexity for one word sequence (i.e. a sentences) $w_{1:L}$ can be calculated as follows, L is the length of the word sequences:

$$\log_2 \mathcal{PPL}(w_{1:L}|\mathbf{I}) = -\frac{1}{L} \sum_{n=1}^L \log_2 P(w_n|w_{1:n-1}, \mathbf{I}) \quad (5)$$

where $\mathcal{PPL}(w_{1:L}|\mathbf{I})$ denotes the perplexity of the sentence $w_{1:L}$ given the image \mathbf{I} . $P(w_n|w_{1:n-1}, \mathbf{I})$ is the probability of the current word w_n given \mathbf{I} and previous words $w_{1:n-1}$. It corresponds to the feature vector of the SoftMax layer of our modal.

The cost function of our model is the average of the logarithm of the perplexity for all the sentences in the training set:

$$\mathcal{C} = -\frac{1}{N} \sum_{i=1}^N \log_2 \mathcal{PPL}(w_{1:L}^{(i)}|\mathbf{I}^{(i)}) \quad (6)$$

where N is the number of sentences in the training set. It is equivalent to the reciprocal of the geometric mean of the probability for the model to generate the training sentences. Our training objective is to minimize this cost function, which is equivalent to maximize the probability of the model to generate the sentences in the training set given their corresponding images. The cost function is derivable thus we can use backpropagation to learn the model parameters.

5 Learning of Image and Sentence Features

The architecture of our model allows the gradients from the loss function to be backpropagated to both the language modeling part (i.e. the word embedding, hidden and the recurrent layers) as well as the image part (e.g. the AlexNet [17]).

For the language modeling part, as mentioned above, we randomly initialize the language modeling layers and learn their parameters. For the image part, we connect the seventh layer of a pre-trained Convolutional Neural Network [17, 3] (also denoted as AlexNet). The same features extract from the seventh layer of AlexNet (also denoted as decaf features [3]) was widely used by previous multimodal methods [16, 6, 15, 29]. In a most recent work [15], the same image features combined with the state-of-the-art detection framework of Region-CNN [7] were used. Their experiments showed that using this detection framework will indeed increase the performance. In the experiments, we will show that our method performs much better than [15] when the same image features are used, and even better than their results of more sophisticated detection features in terms of many evaluation metrics.

We can update the AlexNet according to the gradient backpropagated from the multimodal layer. In this paper, we fix the image features and the deep CNN network in the training stage due to the luck of the data (The datasets we used in the experiment have less than 30K images). In the future work, we will try our method on large dataset and finetune the parameters of the deep CNN network in the training stage.

6 Sentence Generation, Image and Sentence Retrieval

After training of the m-RNN model, we can use the model for the tasks of sentence generation, image retrieval using sentences and sentence retrieval using images.

The sentence generation process is straight forward. Start from the start sign “##START##” or any length of the context words (e.g. we can give the first words in the reference sentences), our model can calculate the probability distribution of the next word: $P(w|w_{1:n-1}, \mathbf{I})$. Then we can sample from this probability distribution to pick the next word. In practice, we find that picking the word with the maximum probability will perform slightly better than sampling. After that, we input the picked word to the model and sample the next word. This process repeated until we have the end sign “##END##”.

For the retrieval tasks, we can use our model to calculate the perplexity of generating a sentence given an image. The perplexity can be treated as an affinity measurement between sentences and images. For the image retrieval task, we just need to retrieve the images that generate the minimum perplexity with the sentence query.

The sentence retrieval tasks is trickier because there might be some sentences that has high probability to any image query. Instead of looking at the perplexity or the probability of the sentences given the query image, we need to use the normalized probability for each sentence: $P(w_{1:L}|\mathbf{I}) / \sum_{\mathbf{I}'} P(w_{1:L}|\mathbf{I}')$ where \mathbf{I}' are images sampled from the training set, $P(w_{1:L}|\mathbf{I}) = \mathcal{PPL}(w_{1:L}|\mathbf{I})^{-L}$.

7 Experiments

7.1 Datasets

We test our method on three benchmark datasets with sentence level annotations: IAPR TC-12 [8], flickr 8K [27], and flickr 30K [13].

Here are some statistics and our experimental settings for the three datasets:

IAPR TC-12 Benchmark This dataset consists of around 20,000 images taken from locations around the world. This includes images of different sports and actions, people, animals, cities, landscapes, and so on. For each image, they provide at least one sentences annotations. On average, there are about 1.7 sentences annotations for one image. We adopt the publicly available separation of training and testing set as previous works [9, 16]. There are 17,665 images for training and 1962 images for testing.

Flickr8K Benchmark This dataset consists of 8,000 images extracted from Flickr. For each image, they provide five sentences annotations. The grammar for the annotations of this dataset is simpler than those of the IAPR TC-12 dataset. We adopt the standard separation of training, validation and testing set which is provided by the dataset. There are 6,000 images for training, 1,000 images for validation and 1,000 images for testing.

Flickr30K Benchmark This dataset is a recent extension of Flickr8K. It consists of 158,915 crowd-sourced captions describing 31,783 images. So for each image, they also provide five sentences annotations. The grammar and style for the annotations of this dataset is similar to Flickr8K. We follow the previous work [15] which used 1,000 images for testing. This dataset, as well as the Flickr8K dataset, is mainly used for the image-sentence retrieval tasks and there is not public available results of methods for generating novel sentence descriptions.

7.2 Experiments settings

Our model can be used for three tasks: 1) Sentences generation; 2) Sentence retrieval (retrieval top relevant sentences to the given image); 3) Image retrieval (retrieval top relevant images to the given sentence);

7.2.1 Evaluation metrics for sentence generation

Following previous works, we use sentence perplexity and BLEU score [26, 20] as the evaluation metrics. BLEU score was originally designed for automatically machine translation where the task is to give a score to a translated sentences given several references sentences. We can treat the sentence generation task as the "translation" of the content of image to sentences. The drawback of using BLEU in our task is that for some images, the reference sentences might not contains all the elements and content in the image and the BLEU score might penalize the arguably correct generated sentences, though it remains as the standard evaluation metric for sentence generation methods for image. To conduct a fair comparison, we adopt the same sentence generation steps and experiment settings as [16], and generate as many words as there are in the reference sentences. Note that our model actually do not need to know the length of the reference sentence because we add a end sign "##END##" at the end of every training sentences and we can stop the generation process when our model outputs the word "##END##".

7.2.2 Evaluation metrics for sentence retrieval and image retrieval

For Flickr8K and Flickr30K datasets, we adopted the same evaluation metrics as previous works [29, 6, 15] for both the tasks of sentences retrieval and image retrieval. They used R@K (K = 1, 5, 10), which is the recall rate of the first groundtruth sentences (sentence retrieval task) or images (image retrieval task) as the measurements. Higher R@K usually means better retrieval performance of different methods. Since we care most of the top retrieved results, the R@K with smaller K is more important than those with larger K. In addition to R@K, they used the Med r, which is the median rank of the first groundtruth sentences (sentence retrieval task) or images (image retrieval task). Lower Med r usually means better performance.

For IAPR TC-12 datasets, we adopt exactly the same evaluation metrics as [16], which plotted the mean number of matches of the retrieved groundtruth sentences or images with respect to the percentage of the retrieved sentences or images for the testing set. For sentences retrieval task, [16] used a shortlist of 100 images which are the nearest neighbors of specific testing image in the image feature space. This shortlist makes the task harder because similar images might have similar descriptions and it is often harder to find the subtle difference among the sentences and pick the

most suitable one. Although there is no published R@K score and Med r score for this dataset as the best of our knowledge, we also report these metrics of our method for future comparison.

7.3 Results on IAPR TC-12

	PERP	B-1	B-2	B-3
BACK-OFF GT2	54.5	0.323	0.145	0.059
BACK-OFF GT3	55.6	0.312	0.131	0.059
LBL [25]	20.1	0.327	0.144	0.068
MLBL-B-DeCAF [16]	24.7	0.373	0.187	0.098
MLBL-F-DeCAF [16]	21.8	0.361	0.176	0.092
Gupta et al. [11]	/	0.15	0.06	0.01
Gupta & Mannem [10]	/	0.33	0.18	0.07
Ours-RNN-Base	7.77	0.3134	0.1168	0.0803
Ours-m-RNN	6.92	0.3951	0.1828	0.1311

Table 1: Results of generated sentences in the iaprtc-12 dataset.

The results of generated sentences is shown in Table 1. BACK-OFF GT2 and GT3 are n-grams methods with Katz backoff and Good-Turing discounting [2, 16]. Ours-RNN-Base has the same architecture with our m-RNN model except that we will not input the image features to the network. It serves as a baseline for our m-RNN model.

To conduct a fair comparison, we followed the same experimental settings of [16], includes the context length to calculate the BLEU score and perplexity. Please note that the perplexity is calculated according to the conditional probability of the words given all of its previous reference words in the sentences. Therefore, a strong language model that successfully captures the grammar of sentences can have a low perplexity without the image content. Perplexity does not directly correlate to the BLEU score where we need to sample the words from the probability distribution generated by the model. For example, although for this dataset, our baseline method of RNN can generate a very low perplexity, it failed to generated sentences with high quality since its BLEU score is not very high. From this perspective, the BLEU score is a better measurement for the generating sentences.

From the experiments, we can see that our m-RNN model performs much better than our baseline RNN model in terms of both perplexity and BLEU score. It also outperforms the state-of-the-art methods in terms of perplexity, B-1, B-3, and a comparable result for B-2.

For retrieval tasks, as mentioned in Section 7.2.2, we draw a recall accuracy curve with respect to the percentage of retrieved images (Text to Image) or retrieved sentences (Image to Text) shown in Figure 3. For sentence retrieval task, we used a shortlist of 100 images as the three comparing methods [16]. The first method, bowdecaf, is a strong image based bag-of-words baseline. The second and the third models are all multimodal deep models. Our m-RNN model outperforms these three methods by a large margin.

Since there are no publicly available results of R@K and median rank in this dataset, we report R@K scores of our method in table 2 for future comparisons.

	Sentence Retrival (Image to Text)				Image Retrival (Text to Image)			
	R@1	R@5	R@10	Med r	R@1	R@5	R@10	Med r
Ours-m-RNN	20.9	43.8	54.4	8	13.2	31.2	40.8	21

Table 2: Results of R@K and median rank (Med r) for iaprtc-12 dataset.

7.4 Results on flickr8K

This dataset was widely used for a benchmark dataset of Text to Image retrieval and Image to Text retrieval. There are no publicly available methods that reports the statistics of generated sentences.

We first show the R@K evaluation metric in Table 3. Our model outperforms the state-of-the-art methods (i.e Socher-decaf, DeVISE-decaf, DeepFE-decaf) by a large margin when using the

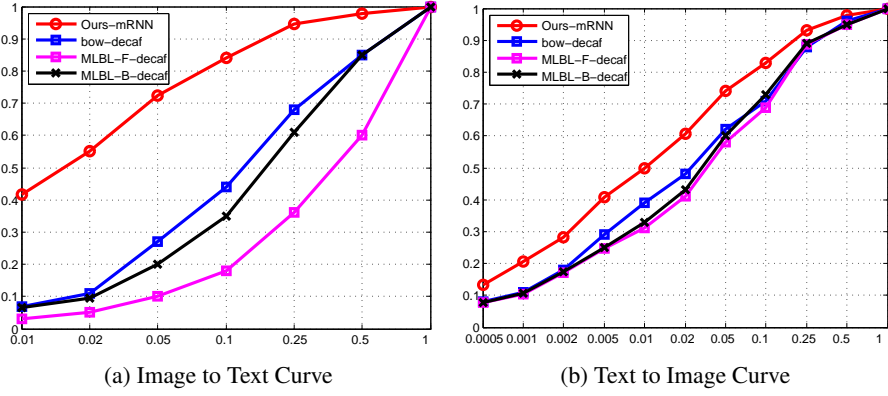


Figure 3: Retrieval recall curve for (a). Sentence retrieval task (Image to Text) (b). Image retrieval task (Text to Image) of iaprtc-12 dataset.

same image features (i.e. decaf features). In a recent work [15], they showed that using more sophisticated image features (e.g. decaf feature combined with detection results), will increase the performance. We also list the result of the methods using such features in Table 3. Socher-avg-rcnn and DeVISE-avg-rcnn used features of the average CNN activation of all objects above a detection confidence threshold [15]. DeepFE-rcnn used a image feature that further utilizes the RCNN detection algorithm. Socher-avg-rcnn and DeepFE-rcnn all shows better results than their original version of Socher-decaf and DeepFE-decaf. From the table, we can see that our method even performs better than these methods in most of the evaluation metrics. We will develop our framework using the detection algorithm in the future work.

	Sentence Retrieval (Image to Text)				Image Retrieval (Text to Image)			
	R@1	R@5	R@10	Med r	R@1	R@5	R@10	Med r
Random	0.1	0.5	1.0	631	0.1	0.5	1.0	500
Socher-decaf [29]	4.5	18.0	28.6	32	6.1	18.5	29.0	29
Socher-avg-rcnn [29]	6.0	22.7	34.0	23	6.6	21.6	31.7	25
DeViSE-avg-rcnn [6]	4.8	16.5	27.3	28	5.9	20.1	29.6	29
DeepFE-decaf [15]	5.9	19.2	27.3	34	5.2	17.6	26.5	32
DeepFE-rcnn [15]	12.6	32.9	44.0	14	9.7	29.6	42.5	15
Ours-m-RNN-decaf	14.5	37.2	48.5	11	11.5	31.0	42.4	15

Table 3: Results of R@K and median rank (Med r) for flickr8K dataset.

We also reports the results of generated sentences in Table 4. There is no publicly available algorithm that reported results on this dataset. We compared our m-RNN model with the Ours-RNN-Base model. The m-RNN model performs much better than the baseline both in terms of the perplexity and BLEU scores.

	PERP	B-1	B-2	B-3
Ours-RNN-Base	30.39	0.4383	0.1849	0.1339
Ours-m-RNN	24.39	0.5778	0.2751	0.2307

Table 4: Results of generated sentences in the flickr8K dataset.

7.5 Results on flickr30K

This dataset is a new dataset and there are only a few methods report their retrieval results on it. We first show the R@K evaluation metric in Table 5. Our method outperforms the state-of-the-art methods in most of the evaluation metrics.

In addition, no publicly available methods reported the results of the sentence generation task as the best of our knowledge. We report the results of generated sentences in Table 6 with a comparison of our RNN baseline.

	Sentence Retrieval (Image to Text)				Image Retrieval (Text to Image)			
	R@1	R@5	R@10	Med r	R@1	R@5	R@10	Med r
Random	0.1	0.6	1.1	631	0.1	0.5	1.0	500
DeViSE-avg-rcnn [6]	4.8	16.5	27.3	28	5.9	20.1	29.6	29
DeepFE-rcnn [15]	16.4	40.2	54.7	8	10.3	31.4	44.5	13
Ours-m-RNN-decaf	18.4	40.2	50.9	10	12.6	31.2	41.5	16

Table 5: Results of R@K and median rank (Med r) for flickr30K dataset.

	PERP	B-1	B-2	B-3
Ours-RNN-Base	43.96	0.4699	0.1964	0.1252
Ours-m-RNN	35.11	0.5479	0.2392	0.1952

Table 6: Results of generated sentences in the flickr8K dataset.

8 Conclusion

We propose a multimodal Recurrent Neural Network framework that perform at the state-of-the-art in the tasks of sentence generation, sentence retrieval given query image and image retrieval given query sentence. In the future work, we will try to incorporate more powerful language models and image feature extraction networks into our framework and try it on larger datasets.

References

- [1] K. Barnard, P. Duygulu, D. Forsyth, N. De Freitas, D. M. Blei, and M. I. Jordan. Matching words and pictures. *JMLR*, 3:1107–1135, 2003.
- [2] S. F. Chen and R. Rosenfeld. A survey of smoothing techniques for me models. *TSAP*, 8(1):37–50, 2000.
- [3] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. *arXiv preprint arXiv:1310.1531*, 2013.
- [4] J. L. Elman. Finding structure in time. *Cognitive science*, 14(2):179–211, 1990.
- [5] A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth. Every picture tells a story: Generating sentences from images. In *ECCV*, pages 15–29, 2010.
- [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.
- [7] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014.
- [8] M. Grubinger, P. Clough, H. Müller, and T. Deselaers. The iapr tc-12 benchmark: A new evaluation resource for visual information systems. In *International Workshop OntoImage*, pages 13–23, 2006.
- [9] M. Guillaumin, J. Verbeek, and C. Schmid. Multiple instance metric learning from automatically labeled bags of faces. In *ECCV*, pages 634–647, 2010.
- [10] A. Gupta and P. Mannem. From image annotation to image description. In *ICONIP*, pages 196–204, 2012.
- [11] A. Gupta, Y. Verma, and C. Jawahar. Choosing linguistics over vision to describe images. In *AAAI*, 2012.
- [12] M. Hodosh, P. Young, and J. Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. *JAIR*, 47:853–899, 2013.
- [13] P. Y. A. L. M. Hodosh and J. Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions.
- [14] Y. Jia, M. Salzmann, and T. Darrell. Learning cross-modality similarity for multinomial data. In *ICCV*, pages 2407–2414, 2011.
- [15] A. Karpathy, A. Joulin, and L. Fei-Fei. Deep fragment embeddings for bidirectional image sentence mapping. In *arXiv:1406.5679*, 2014.
- [16] R. Kiros, R. Zemel, and R. Salakhutdinov. Multimodal neural language models. In *ICML*, 2014.
- [17] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1097–1105, 2012.

- 486 [18] G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg. Baby talk: Understanding
487 and generating image descriptions. In *CVPR*, 2011.
- 488 [19] Y. A. LeCun, L. Bottou, G. B. Orr, and K.-R. Müller. Efficient backprop. In *Neural networks: Tricks of*
489 *the trade*, pages 9–48. Springer, 2012.
- 490 [20] C.-Y. Lin and F. J. Och. Automatic evaluation of machine translation quality using longest common
491 subsequence and skip-bigram statistics. In *ACL*, page 605, 2004.
- 492 [21] T. Mikolov, M. Karafiát, L. Burget, J. Cernocký, and S. Khudanpur. Recurrent neural network based
493 language model. In *INTERSPEECH*, pages 1045–1048, 2010.
- 494 [22] T. Mikolov, S. Kombrink, L. Burget, J. Cernocký, and S. Khudanpur. Extensions of recurrent neural
495 network language model. In *ICASSP*, pages 5528–5531, 2011.
- 496 [23] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and
497 phrases and their compositionality. In *NIPS*, pages 3111–3119, 2013.
- 498 [24] M. Mitchell, X. Han, J. Dodge, A. Mensch, A. Goyal, A. Berg, K. Yamaguchi, T. Berg, K. Stratos, and
499 H. Daumé III. Midge: Generating image descriptions from computer vision detections. In *EACL*, pages
500 747–756, 2012.
- 501 [25] A. Mnih and G. Hinton. Three new graphical models for statistical language modelling. In *ICML*, pages
502 641–648. ACM, 2007.
- 503 [26] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: a method for automatic evaluation of machine
504 translation. In *ACL*, pages 311–318, 2002.
- 505 [27] C. Rashtchian, P. Young, M. Hodosh, and J. Hockenmaier. Collecting image annotations using amazon’s
506 mechanical turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language*
507 *Data with Amazon’s Mechanical Turk*, pages 139–147. Association for Computational Linguistics, 2010.
- 508 [28] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning representations by back-propagating errors.
509 *Cognitive modeling*, 1988.
- 510 [29] R. Socher, Q. Le, C. Manning, and A. Ng. Grounded compositional semantics for finding and describing
511 images with sentences. In *TACL*, 2014.
- 512 [30] N. Srivastava and R. Salakhutdinov. Multimodal learning with deep boltzmann machines. In *NIPS*, pages
513 2222–2230, 2012.
- 514
- 515
- 516
- 517
- 518
- 519
- 520
- 521
- 522
- 523
- 524
- 525
- 526
- 527
- 528
- 529
- 530
- 531
- 532
- 533
- 534
- 535
- 536
- 537
- 538
- 539