Learning a Recurrent Visual Representation for Image Caption Generation

Xinlei Chen
Carnegie Mellon University
xinleic@cs.cmu.edu

C. Lawrence Zitnick
Microsoft Research, Redmond
larryz@microsoft.com

Abstract

In this paper we explore the bi-directional mapping between images and their sentence-based descriptions. We propose learning this mapping using a recurrent neural network. Unlike previous approaches that map both sentences and images to a common embedding, we enable the generation of novel sentences given an image. Using the same model, we can also reconstruct the visual features associated with an image given its visual description. We use a novel recurrent visual memory that automatically learns to remember long-term visual concepts to aid in both sentence generation and visual feature reconstruction. We evaluate our approach on several tasks. These include sentence generation, sentence retrieval and image retrieval. State-of-the-art results are shown for the task of generating novel image descriptions. When compared to human generated captions, our automatically generated captions are preferred by humans over 19.8% *of the time. Results are better than or comparable to state-of-the-art results on the image and sentence retrieval tasks for methods using similar visual features.*

1. Introduction

A good image description is often said to "paint a picture in your mind's eye." The creation of a mental image may play a significant role in sentence comprehension in humans [15]. In fact, it is often this mental image that is remembered long after the exact sentence is forgotten [29, 21]. What role should visual memory play in computer vision algorithms that comprehend and generate image descriptions?

Recently, several papers have explored learning joint feature spaces for images and their descriptions [13, 32, 16]. These approaches project image features and sentence features into a common space, which may be used for image search or for ranking image captions. Various approaches were used to learn the projection, including Kernel Canonical Correlation Analysis (KCCA) [13], recursive neural networks [32], or deep neural networks [16]. While these approaches project both semantics and visual features to a common embedding, they are not able to perform the inverse projection. That is, they cannot generate novel sentences or visual depictions from the embedding.

In this paper, we propose a bi-directional representation capable of generating both novel descriptions from images and visual representations from descriptions. Critical to both of these tasks is a novel representation that dynamically captures the visual aspects of the scene that have already been described. That is, as a word is generated or read the visual representation is updated to reflect the new information contained in the word. We accomplish this using Recurrent Neural Networks (RNNs) [6, 24, 27]. One long-standing problem of RNNs is their weakness in remembering concepts after a few iterations of recurrence. For instance RNN language models often find difficultly in learning long distance relations [3, 24] without specialized gating units [12]. During sentence generation, our novel dynamically updated visual representation acts as a long-term memory of the concepts that have already been mentioned. This allows the network to automatically pick salient concepts to convey that have yet to be spoken. As we demonstrate, the same representation may be used to create a visual representation of a written description.

We demonstrate our method on numerous datasets. These include the PASCAL sentence dataset [31], Flickr 8K [31], Flickr 30K [31], and the Microsoft COCO dataset [22]. When generating novel image descriptions, we demonstrate state-of-the-art results as measured by both BLEU [30] and METEOR [1] on PASCAL 1K. Surprisingly, we achieve performance only slightly below humans as measured by BLEU and METEOR on the MS COCO dataset. Qualitative results are shown for the generation of novel image captions. We also evaluate the bi-directional ability of our algorithm on both the image and sentence retrieval tasks. Since this does not require the ability to generate novel sentences, numerous previous papers have evaluated on this task. We show results that are better or comparable to previous state-of-the-art results using similar visual features.

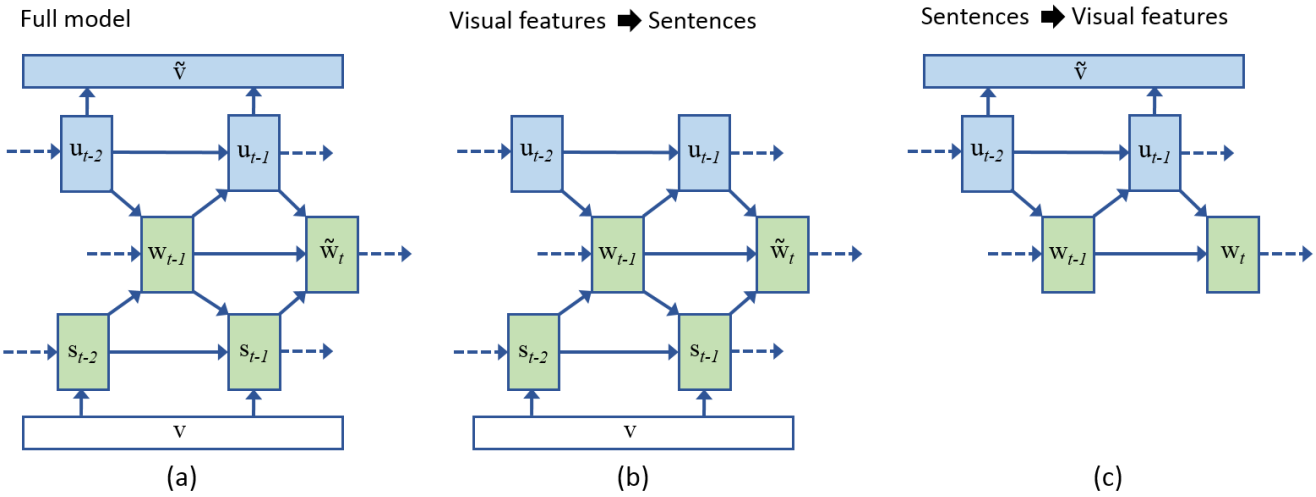

Figure 1. Illustration of our model. (a) shows the full model used for training. (b) and (c) show the parts of the model needed for generating sentences from visual features and generating visual features from sentences respectively.

2. Related work

The task of building a visual memory lies at the heart of two long-standing AI-hard problems: grounding natural language symbols to the physical world and semantically understanding the content of an image. Whereas learning the mapping between image patches and single text labels remains a popular topic in computer vision [18, 8, 9], there is a growing interest in using entire sentence descriptions together with pixels to learn joint embeddings [13, 32, 16, 10]. Viewing corresponding text and images as correlated, KCCA [13] is a natural option to discover the shared features spaces. However, given the highly non-linear mapping between the two, finding a generic distance metric based on shallow representations can be extremely difficult. Recent papers seek better objective functions that directly optimize the ranking [13], or directly adopts pre-trained representations [32] to simplify the learning, or a combination of the two [16, 10].

With a good distance metric, it is possible to perform tasks like bi-directional image-sentence retrieval. However, in many scenarios it is also desired to generate novel image descriptions and to hallucinate a scene given a sentence description. Numerous papers have explored the area of generating novel image descriptions [7, 36, 19, 37, 28, 11, 20, 17]. These papers use various approaches to generate text, such as using pre-trained object detectors with template-based sentence generation [36, 7, 19]. Retrieved sentences may be combined to form novel descriptions [20]. Recently, purely statistical models have been used to generate sentences based on sampling [17] or recurrent neural networks [23]. While [23] also uses a RNN, their model is significantly different from our model. Specifically their RNN does not attempt to reconstruct the visual features, and is more similar to the contextual RNN of [27]. For the synthesizing of images from sentences, the recent paper by Zitnick *et al*. [38] uses abstract clip art images to learn the visual interpretation of sentences. Relation tuples are extracted from the sentences and a conditional random field is used to model the visual scene.

There are numerous papers using recurrent neural networks for language modeling [2, 24, 27, 17]. We build most directly on top of [2, 24, 27] that use RNNs to learn word context. Several models use other sources of contextual information to help inform the language model [27, 17]. Despite its success, RNNs still have difficulty capturing long-range relationships in sequential modeling [3]. One solution is Long Short-Term Memory (LSTM) networks [12, 33, 17], which use "gates" to control gradient back-propagation explicitly and allow for the learning of long-term interactions. However, the main focus of this paper is to show that the hidden layers learned by "translating" between multiple modalities can already discover rich structures in the data and learn long distance relations in an automatic, data-driven manner.

3. Approach

In this section we describe our approach using recurrent neural networks. Our goals are twofold. First, we want to be able to generate sentences given a set of visual observations or features. Specifically, we want to compute the probability of a word w_t being generated at time t given the set of previously generated words $W_{t-1} = w_1, \ldots, w_{t-1}$ and the observed visual features V. Second, we want to enable the capability of computing the likelihood of the visual features V given a set of spoken or read words W_t for generating visual representations of the scene or for performing image search. To accomplish both of these tasks we introduce a

set of latent variables U_{t-1} that encodes the visual interpretation of the previously generated or read words W_{t-1}. As we demonstrate later, the latent variables U play the critical role of acting as a long-term visual memory of the words that have been previously generated or read.

Using U, our goal is to compute $P(w_t|V, W_{t-1}, U_{t-1})$ and $P(V|W_{t-1}, U_{t-1})$. Combining these two likelihoods together our global objective is to maximize,

$$P(w_t, V|W_{t-1}, U_{t-1}) = P(w_t|V, W_{t-1}, U_{t-1})P(V|W_{t-1}, U_{t-1}). \quad (1)$$

That is, we want to maximize the likelihood of the word w_t and the observed visual features V given the previous words and their visual interpretation. Note that in previous papers [27, 23] the objective was only to compute $P(w_t|V, W_{t-1})$ and not $P(V|W_{t-1})$.

3.1. Model structure

Our recurrent neural network model structure builds on the prior models proposed by [24, 27]. Mikolov [24] proposed a RNN language model shown by the green boxes in Figure 1(a). The word at time t is represented by a vector $\mathbf{w}_t$ using a "one hot" representation. That is, $\mathbf{w}_t$ is the same size as the word vocabulary with each entry having a value of 0 or 1 depending on whether the word was used. The output $\tilde{\mathbf{w}}_t$ contains the likelihood of generating each word. The recurrent hidden state $\mathbf{s}$ provides context based on the previous words. However, $\mathbf{s}$ typically only models short-range interactions due to the problem of vanishing gradients [3, 24]. This simple, yet effective language model was shown to provide a useful continuous word embedding for a variety of applications [25].

Following [24], Mikolov *et al*. [27] added an input layer $\mathbf{v}$ to the RNN shown by the white box in Figure 1. This layer may represent a variety of information, such as topic models or parts of speech [27]. In our application, $\mathbf{v}$ represents the set of observed visual features. We assume the visual features $\mathbf{v}$ are constant. These visual features help inform the selection of words. For instance, if a cat was detected, the word "cat" is more likely to be spoken. Note that unlike [27], it is not necessary to directly connect $\mathbf{v}$ to $\tilde{\mathbf{w}}$, since $\mathbf{v}$ is static for our application. In [27] $\mathbf{v}$ represented dynamic information such as parts of speech for which $\tilde{\mathbf{w}}$ needed direct access. We also found that only connecting $\mathbf{v}$ to half of the $\mathbf{s}$ units provided better results, since it allowed different units to specialize on modeling either text or visual features.

The main contribution of this paper is the addition of the recurrent visual hidden layer $\mathbf{u}$, blue boxes in Figure 1(a). The recurrent layer $\mathbf{u}$ attempts to reconstruct the visual features $\mathbf{v}$ from the previous words, *i.e*. $\tilde{\mathbf{v}} \approx \mathbf{v}$. The visual hidden layer is also used by $\tilde{\mathbf{w}}_t$ to help in predicting the next word. That is, the network can compare its visual memory of what it has already said $\mathbf{u}$ to what it currently observes $\mathbf{v}$ to predict what to say next. At the beginning of the sentence, $\mathbf{u}$ represents the prior probability of the visual features. As more words are observed, the visual feature likelihoods are updated to reflect the words' visual interpretation. For instance, if the word "sink" is generated, the visual feature corresponding to sink will increase. Other features that correspond to stove or refrigerator might increase as well, since they are highly correlated with sink.

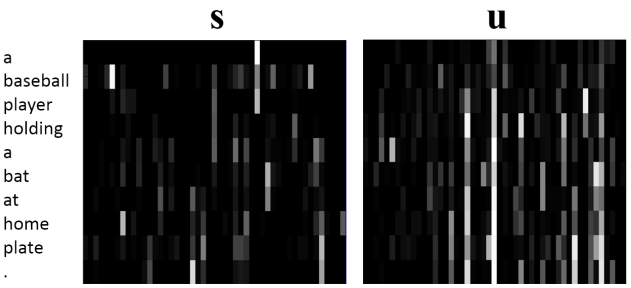

Figure 2. Illustration of the hidden units $\mathbf{s}$ and $\mathbf{u}$ activations through time (vertical axis). Notice that the visual hidden units $\mathbf{u}$ exhibit long-term memory through the temporal stability of some units, where the hidden units $\mathbf{s}$ change significantly each time step.

A critical property of the recurrent visual features $\mathbf{u}$ is their ability to remember visual concepts over the long term. The property arises from the model structure. Intuitively, one may expect the visual features shouldn't be estimated until the sentence is finished. That is, $\mathbf{u}$ should not be used to estimate $\mathbf{v}$ until $\mathbf{w}_t$ generates the end of sentence token. However, in our model we force $\mathbf{u}$ to estimate $\mathbf{v}$ at every time step to help in remembering visual concepts. For instance, if the word "cat" is generated, $\mathbf{u}_t$ will increase the likelihood of the visual feature corresponding to cat. Assuming the "cat" visual feature in $\mathbf{v}$ is active, the network will receive positive reinforcement to propagate $\mathbf{u}$'s memory of "cat" from one time instance to the next. Figure 2 shows an illustrative example of the hidden units $\mathbf{s}$ and $\mathbf{u}$. As can be observed, some visual hidden units $\mathbf{u}$ exhibit longer temporal stability.

Note that the same network structure can predict visual features from sentences or generate sentences from visual features. For generating sentences (Fig. 1(b)), $\mathbf{v}$ is known and $\tilde{\mathbf{v}}$ may be ignored. For predicting visual features from sentences (Fig. 1(c)), $\mathbf{w}$ is known, and $\mathbf{s}$ and $\mathbf{v}$ may be ignored. This property arises from the fact that the words units $\mathbf{w}$ separate the model into two halves for predicting words or visual features respectively. Alternatively, if the hidden units $\mathbf{s}$ were connected directly to $\mathbf{u}$, this property would be lost and the network would act as a normal auto-encoder [34].

3.2. Implementation details

In this section we describe the details of our language model and how we learn our network.

3.3. Language Model

Our language model typically has between 3,000 and 20,000 words. While each word may be predicted independently, this approach is computationally expensive. Instead, we adopted the idea of word classing [24] and factorized the distribution into a product of two terms:

$$P(w_t|\cdot) = P(c_t|\cdot) \times P(w_t|c_t, \cdot). \quad (2)$$

$P(w_t|\cdot)$ is the probability of the word, $P(c_t|\cdot)$ is the probability of the class. The class label of the word is computed in an unsupervised manner, grouping words of similar frequencies together. Generally, this approach greatly accelerates the learning process, with little loss of perplexity. The predicted word likelihoods are computed using the standard soft-max function. After each epoch, the perplexity is evaluated on a separate validation set and the learning reduced (cut in half in our experiments) if perplexity does not decrease.

In order to further reduce the perplexity, we combine the RNN model's output with the output from a Maximum Entropy language model [26], simultaneously learned from the training corpus. For all experiments we fix how many words to look back when predicting the next word used by the Maximum Entropy model to three.

For any natural language processing task, pre-processing is crucial to the final performance. For all the sentences, we did the following two steps before feeding them into the RNN model. 1) Use Stanford CoreNLP Tool to tokenize the sentences. 2) Lower case all the letters.

3.4. Learning

For learning we use the Backpropagation Through Time (BPTT) algorithm [35]. Specifically, the network is unrolled for several words and standard backpropagation is applied. Note that we reset the model after an End-of-Sentence (EOS) is encountered, so that prediction does not cross sentence boundaries. As shown to be beneficial in [24], we use online learning for the weights from the recurrent units to the output words. The weights for the rest of the network use a once per sentence batch update. The activations for all units are computed using the sigmoid function $\sigma(z) = 1/(1 + \exp(-z))$ with clipping, except the word predictions that use soft-max. We found that Rectified Linear Units (ReLUs) [18] with unbounded activations were numerically unstable and commonly "blew up" when used in recurrent networks.

We used the open source RNN code of [24] and the Caffe framework [14] to implement our model. A big advantage of combining the two is that we can jointly learn the word and image representations: the error from predicting the words can be directly backpropagated to the image-level features. However, deep convolution neural networks require large amounts of data to train on, but the largest sentence-image dataset has only ∼80K images [22]. Therefore, instead of training from scratch, we choose to fine-tune from the pre-trained 1000-class ImageNet model [4] to avoid potential over-fitting. In all experiments, we used the 4096D 7th full-connected layer output as the visual input $\mathbf{v}$ to our model.

4. Results

In this section we evaluate the effectiveness of our bi-directional RNN model on multiple tasks. We begin by describing the datasets used for training and testing, followed by our baselines. Our first set of evaluations measure our model's ability to generate novel descriptions of images. Since our model is bi-directional, we evaluate its performance on both the sentence retrieval and image retrieval tasks. For addition results please see the supplementary material.

4.1. Datasets

For evaluation we perform experiments on several standard datasets that are used for sentence generation and the sentence-image retrieval task:

PASCAL 1K [31] The dataset contains a subset of images from the PASCAL VOC challenge. For each of the 20 categories, it has a random sample of 50 images with 5 descriptions provided by Amazon's Mechanical Turk (AMT).

Flickr 8K and 30K [31] These datasets consists of 8,000 and 31,783 images collected from Flickr respectively. Most of the images depict humans participating in various activities. Each image is also paired with 5 sentences. These datasets have a standard training, validation, and testing splits.

MS COCO [22] The Microsoft COCO dataset contains 82,783 training images and 40,504 validation images each with ∼5 human generated descriptions. The images are collected from Flickr by searching for common object categories, and typically contain multiple objects with significant contextual information. We downloaded the version which contains ∼40K annotated training images and ∼10K validation images for our experiments.

4.2. RNN Baselines

To gain insight into the various components of our model, we compared our final model with three RNN baselines. For fair comparison, the random seed initialization

	PASCAL			Flickr 8K			Flickr 30K			MS COCO		
	PPL	BLEU	METR	PPL	BLEU	METR	PPL	BLEU	METR	PPL	BLEU	METR
Midge [28]	-	2.89	8.80					-				
Baby Talk [19]	-	0.49	9.69					-				
RNN	36.79	2.79	10.08	21.88	4.86	11.81	26.94	6.29	12.34	18.96	4.63	11.47
RNN+IF	30.04	10.16	16.43	20.43	12.04	17.10	23.74	10.59	15.56	15.39	16.60	19.24
RNN+IF+FT	29.43	10.18	16.45	-	-	-	-	-	-	14.90	16.77	19.41
Our Approach	27.97	10.48	16.69	19.24	14.10	17.97	22.51	12.60	16.42	14.23	18.35	20.04
Our Approach + FT	26.95	10.77	16.87	-	-	-	-	-	-	13.98	18.99	20.42
Human	-	22.07	25.80	-	22.51	26.31	-	19.62	23.76	-	20.19	24.94

Table 1. Results for novel sentence generation for PASCAL 1K, Flickr 8K, FLickr 30K and MS COCO. Results are measured using perplexity (PPL), BLEU (%) [30] and METEOR (METR, %) [1]. When available results for Midge [28] and BabyTalk [19] are provided. Human agreement scores are shown in the last row. See the text for more details.

	Sentence Retrieval				Image Retrieval			
	R@1	R@5	R@10	Mean r	R@1	R@5	R@10	Mean r
Random Ranking	4.0	9.0	12.0	71.0	1.6	5.2	10.6	50.0
DeViSE [8]	17.0	57.0	68.0	11.9	21.6	54.6	72.4	9.5
SDT-RNN [32]	25.0	56.0	70.0	13.4	35.4	65.2	84.4	7.0
DeepFE [16]	39.0	68.0	79.0	10.5	23.6	65.2	79.8	7.6
RNN+IF	31.0	68.0	87.0	6.0	27.2	65.4	79.8	7.0
Our Approach (T)	25.0	71.0	86.0	5.4	28.0	65.4	82.2	6.8
Our Approach (T+I)	30.0	75.0	87.0	5.0	28.0	67.4	83.4	6.2

Table 2. PASCAL image and sentence retrieval experiments. The protocol of [32] was followed. Results are shown for recall at 1, 5, and 10 recalled items and the mean ranking (Mean r) of ground truth items.

was fixed for all experiments. The the hidden layers $\mathbf{s}$ and $\mathbf{u}$ sizes are fixed to 100. We tried increasing the number of hidden units, but results did not improve. For small datasets, more units can lead to overfitting.

RNN based Language Model (RNN) This is the basic RNN language model developed by [24], which has no input visual features.

RNN with Image Features (RNN+IF) This is an RNN model with image features feeding into the hidden layer inspired by [27]. As described in Section 3 $\mathbf{v}$ is only connected to $\mathbf{s}$ and not $\tilde{\mathbf{w}}$. For the visual features $\mathbf{v}$ we used the 4096D 7th Layer output of BVLC reference Net [14] after ReLUs. This network is trained on the ImageNet 1000-way classification task [4]. We experimented with other layers (5th and 6th) but they do not perform as well.

RNN with Image Features Fine-Tuned (RNN+FT) This model has the same architecture as RNN+IF, but the error is back-propagated to the Convolution Neural Network [9]. The CNN is initialized with the weights from the BVLC reference net. The RNN is initialized with the the pre-trained RNN language model. That is, the only randomly initialized weights are the ones from visual features $\mathbf{v}$ to hidden layers $\mathbf{s}$. If the RNN is not pre-trained we found the initial gradients to be too noisy for the CNN. If the weights from $\mathbf{v}$ to hidden layers $\mathbf{s}$ are also pre-trained the search space becomes too limited. Our current implementation takes $\sim$5 seconds to learn a mini-batch of size 128 on a Tesla K40 GPU. It is also crucial to keep track of the validation error and avoid overfitting. We observed this fine-tuning strategy is particularly helpful for MS COCO, but does not give much performance gain on Flickr Datasets before it overfits. The Flickr datasets may not provide enough training data to avoid overfitting.

After fine-tuning, we fix the image features again and retrain our model on top of it.

4.3. Sentence generation

Our first set of experiments evaluate our model's ability to generate novel sentence descriptions of images. We experiment on all the image-sentence datasets described previously and compare to the RNN baselines and other previous papers [28, 19]. Since PASCAL 1K has a limited amount of training data, we report results trained on MS COCO and tested on PASCAL 1K. We use the standard train-test splits for the Flickr 8K and 30K datasets. For MS COCO we train and validate on the training set ($\sim$37K/$\sim$3K), and test on the validation set, since the testing set is not available. To generate a sentence, we first sample a target sentence length from the multinomial distribution of lengths learned from the training data, then for this fixed length we sample 100 random sentences, and use the one with the lowest loss (negative likelihood, and in case of our model, also reconstruction error) as output.

	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
[32]	4.5	18.0	28.6	32	6.1	18.5	29.0	29
DeViSE [8]	4.8	16.5	27.3	28	5.9	20.1	29.6	29
DeepFE [16]	12.6	32.9	44.0	14	9.7	29.6	42.5	15
DeepFE+DECAF [16]	5.9	19.2	27.3	34	5.2	17.6	26.5	32
RNN+IF	7.2	18.7	28.7	30.5	4.5	15.34	24.0	39
Our Approach (T)	7.6	21.1	31.8	27	5.0	17.6	27.4	33
Our Approach (T+I)	7.7	21.0	31.7	26.6	5.2	17.5	27.9	31
[13]	8.3	21.6	30.3	34	7.6	20.7	30.1	38
RNN+IF	5.5	17.0	27.2	28	5.0	15.0	23.9	39.5
Our Approach (T)	6.0	19.4	31.1	26	5.3	17.5	28.5	33
Our Approach (T+I)	6.2	19.3	32.1	24	5.7	18.1	28.4	31
M-RNN [23]	14.5	37.2	48.5	11	11.5	31.0	42.4	15
RNN+IF	10.4	30.9	44.2	14	10.2	28.0	40.6	16
Our Approach (T)	11.6	33.8	47.3	11.5	11.4	31.8	45.8	12.5
Our Approach (T+I)	11.7	34.8	48.6	11.2	11.4	32.0	46.2	11

Table 3. Flickr 8K Retrieval Experiments. The protocols of [32], [13] and [23] are used respectively in each row. See text for details.

	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 [8]	4.5	18.1	29.2	26	6.7	21.9	32.7	25
DeepFE+FT [16]	16.4	40.2	54.7	8	10.3	31.4	44.5	13
RNN+IF	8.0	19.4	27.6	37	5.1	14.8	22.8	47
Our Approach (T)	9.3	23.8	24.0	28	6.0	17.7	27.0	35
Our Approach (T+I)	9.6	24.0	27.2	25	7.1	17.9	29.0	31
M-RNN [23]	18.4	40.2	50.9	10	12.6	31.2	41.5	16
RNN+IF	9.5	29.3	42.4	15	9.2	27.1	36.6	21
Our Approach (T)	11.9	25.0	47.7	12	12.8	32.9	44.5	13
Our Approach (T+I)	12.1	27.8	47.8	11	12.7	33.1	44.9	12.5

Table 4. Flickr 30K Retrieval Experiments. The protocols of [8] and [23] are used respectively in each row. See text for details.

We choose three automatic metrics for evaluating the quality of the generated sentences, perplexity, BLEU [30] and METEOR [1]. Perplexity measures the likelihood of generating the testing sentence based on the number of bits it would take to encode it. The lower the value the better. BLEU and METEOR were originally designed for automatic machine translation where they rate the quality of a translated sentences given several references sentences. We can treat the sentence generation task as the "translation" of images to sentences. For BLEU, we took the geometric mean of the scores from 1-gram to 4-gram, and used the ground truth length closest to the generated sentence to penalize brevity. For METEOR, we used the latest version[1] (v1.5). For both BLEU and METEOR higher scores are better. For reference, we also report the consistency between human annotators (using 1 sentence as query and the rest as references)[2].

[1] http://www.cs.cmu.edu/~alavie/METEOR/

[2] We used 5 sentences as references for system evaluation, but leave out 4 sentences for human consistency. It is a bit unfair but the difference is usually around 0.01∼0.02.

Results are shown in Table 1. Our approach significantly improves over both Midge [28] and BabyTalk [19] on the PASCAL 1K dataset as measured by BLEU and METEOR. Several qualitative results for the three algorithms are shown in Figure 4. Our approach generally provides more naturally descriptive sentences, such as mentioning an image is black and white, or a bus is a "double decker". Midge's descriptions are often shorter with less detail and BabyTalk provides long, but often redundant descriptions. Results on Flickr 8K and Flickr 30K are also provided.

On the MS COCO dataset that contains more images of high complexity we provide perplexity, BLEU and METEOR scores. Surprisingly our BLEU and METEOR scores (18.99 & 20.42) are just slightly lower than the human scores (20.19 & 24.94). The use of image features (RNN + IF) significantly improves performance over using just an RNN language model. Fine-tuning (FT) and our full approach provide additional improvements for all datasets. For future reference, our final model gives BLEU-1 to BLEU-4 (with penalty) as 60.4%, 26.4%, 12.6% and 6.5%, compared to human consistency 65.9%, 30.5%, 13.6% and

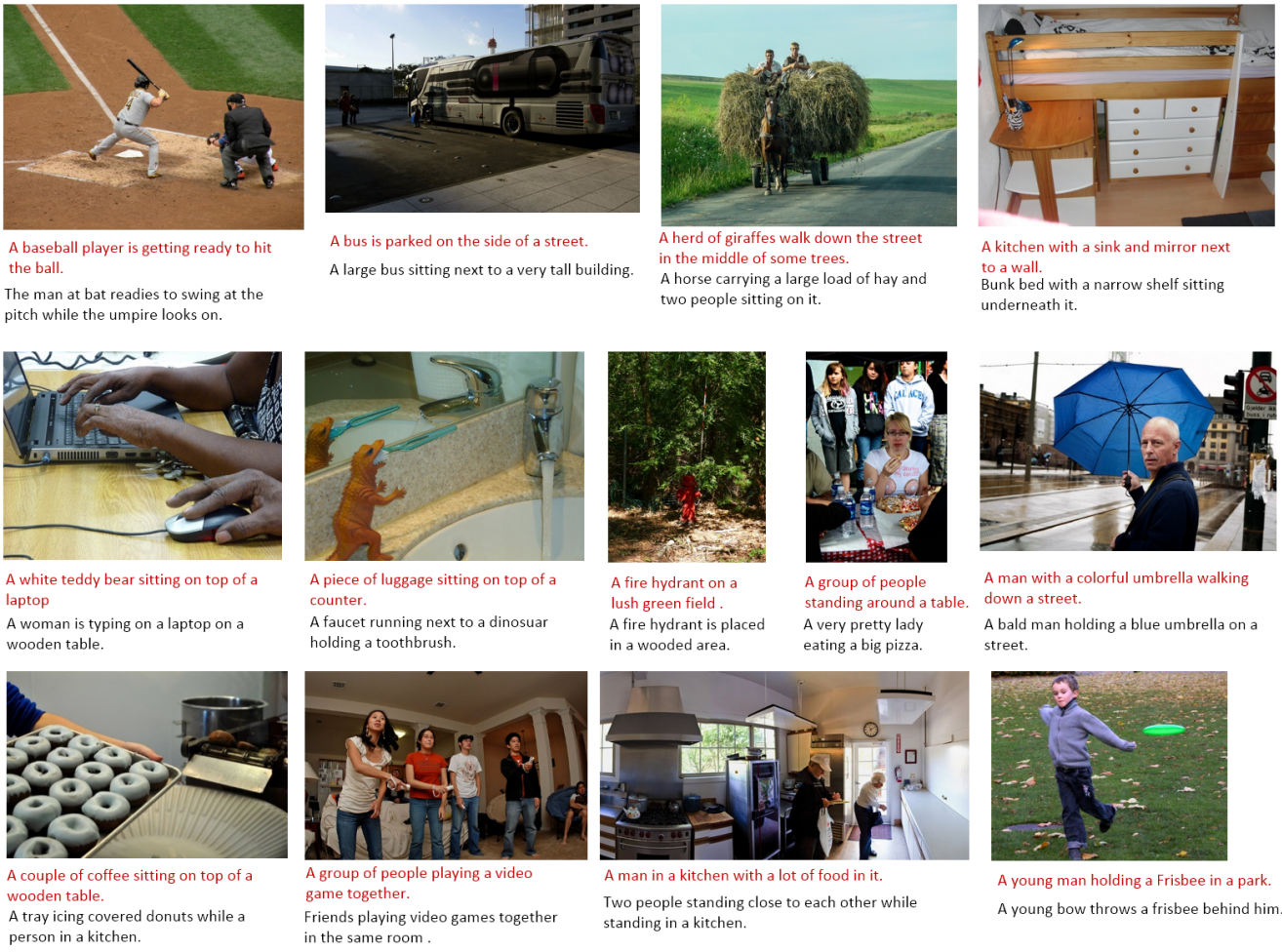

Figure 3. Qualitative results for sentence generation on the MS COCO dataset. Both a generated sentence (red) using (Our Approach + FT) and a human generated caption (black) are shown.

6.0%. Qualitative results for the MS COCO dataset are shown in Figure 3. Note that since our model is trained on MS COCO, the generated sentences are generally better on MS COCO than PASCAL 1K.

It is known that automatic measures are only roughly correlated with human judgment [5], so it is also important to evaluate the generated sentences using human studies. We evaluated 1000 generated sentences on MS COCO by asking human subjects to judge whether it had better, worse or same quality to a human generated ground truth caption. 5 subjects were asked to rate each image, and the majority vote was recorded. In the case of a tie (2-2-1) the two winners each got half of a vote. We find 12.6% and 19.8% prefer our automatically generated captions to the human captions without (Our Approach) and with fine-tuning (Our Approach + FT) respectively. Less than 1% of the subjects rated the captions as the same. This is an impressive result given we only used image-level visual features for the complex images in MS COCO.

4.4. Bi-Directional Retrieval

Our RNN model is bi-directional. That is, it can generate image features from sentences and sentences from image features. To evaluate its ability to do both, we measure its performance on two retrieval tasks. We retrieve images given a sentence description, and we retrieve a description given an image. Since most previous methods are capable of only the retrieval task, this also helps provide experimental comparison.

Following other methods, we adopted two protocols for using multiple image descriptions. The first one is to treat each of the ∼5 sentences individually. In this scenario, the rank of the retrieved ground truth sentences are used for evaluation. In the second case, we treat all the sentences as a single annotation, and concatenate them together for retrieval.

For each retrieval task we have two methods for ranking. First, we may rank based on the likelihood of the sentence

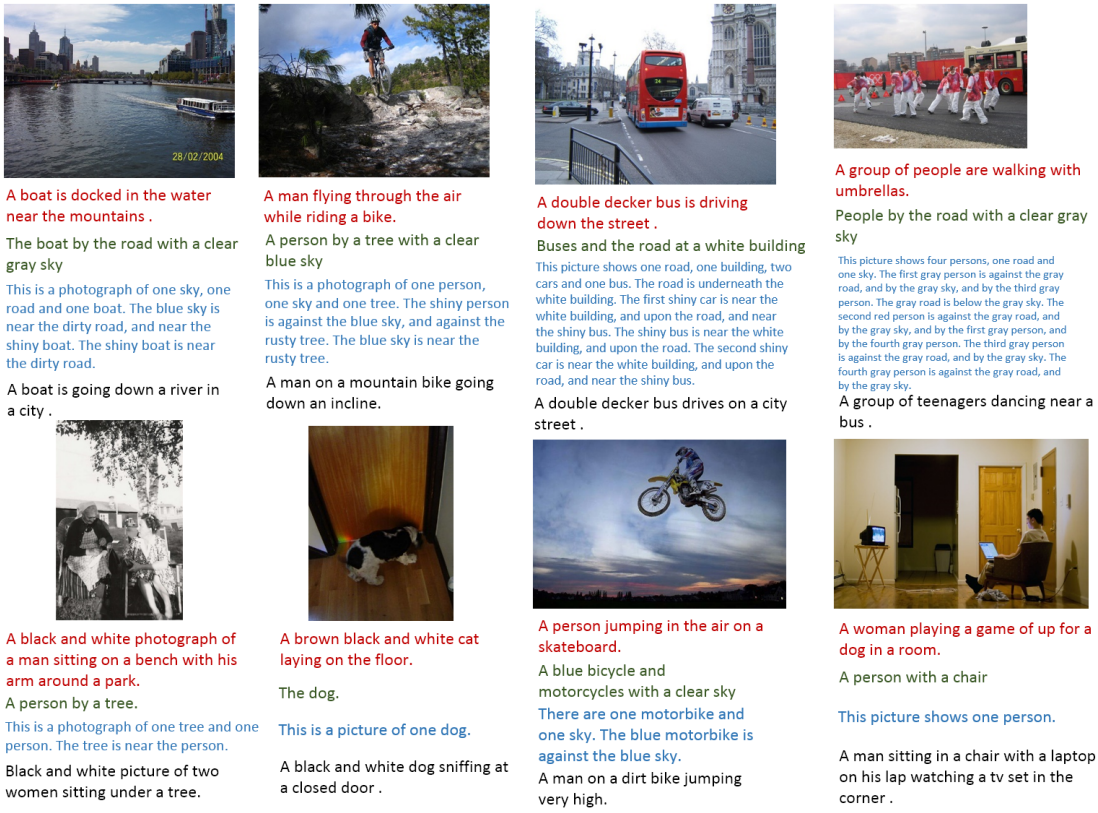

Figure 4. Qualitative results for sentence generation on the PASCAL 1K dataset. Generated sentences are shown for our approach (red), Midge [28] (green) and BabyTalk [19] (blue). For reference, a human generated caption is shown in black.

given the image (T). Since shorter sentences naturally have higher probability of being generated, we followed [23] and normalized the probability by dividing it with the total probability summed over the entire retrieval set. Second, we could rank based on the reconstruction error between the image's visual features $\mathbf{v}$ and their reconstructed visual features $\tilde{\mathbf{v}}$ (I). Due to better performance, we use the average reconstruction error over all time steps rather than just the error at the end of the sentence. In Tables 3, we report retrieval results on using the text likelihood term only (I) and its combination with the visual feature reconstruction error (T+I).

The same evaluation metrics were adopted from previous papers for both the tasks of sentence retrieval and image retrieval. They used R@K (K = 1, 5, 10) as the measurements, which are the recall rates of the (first) ground truth sentences (sentence retrieval task) or images (image retrieval task). Higher R@K corresponds to better retrieval performance. We also report the median/mean rank of the (first) retrieved ground truth sentences or images (Med/Mean r). Lower Med/Mean r implies better performance. For Flickr 8K and 30K several different evaluation methodologies have been proposed. We report three scores for Flickr 8K corresponding to the methodologies proposed by [32], [13] and [23] respectively, and for Flickr 30K [8] and [23].

Measured by Mean r, we achieve state-of-the-art results on PASCAL 1K image and sentence retrieval (Table 2). As shown in Tables 3 and 4, for Flickr 8K and 30K our approach achieves comparable or better results than all methods except for the recently proposed DeepFE [16]. However, DeepFE uses a different set of features based on smaller image regions. If the same features are used (DeepFE+DECAF) as our approach, we achieve better results. We believe these contributions are complementary, and by using better features our approach may also show further improvement. In general ranking based on text and visual features (T + I) outperforms just using text (T). Please see the supplementary material for retrieval results on MS COCO.

5. Discussion

Image captions describe both the objects in the image and their relationships. An area of future work is to examine the sequential exploration of an image and how it relates to image descriptions. Many words correspond to spatial relations that our current model has difficultly in detecting. As demonstrated by the recent paper of [16] better feature localization in the image can greatly improve the performance

of retrieval tasks and similar improvement might be seen in the description generation task.

In conclusion, we describe the first bi-directional model capable of the generating both novel image descriptions and visual features. Unlike many previous approaches using RNNs, our model is capable of learning long-term interactions. This arises from using a recurrent visual memory that learns to reconstruct the visual features as new words are read or generated. We demonstrate state-of-the-art results on the task of sentence generation, image retrieval and sentence retrieval on numerous datasets.

6. Acknowledgements

We thank Hao Fang, Saurabh Gupta, Meg Mitchell, Xiaodong He, Geoff Zweig, John Platt and Piotr Dollar for their thoughtful and insightful discussions in the creation of this paper.

References

[1] S. Banerjee and A. Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, 2005. 1, 5

[2] Y. Bengio, H. Schwenk, J.-S. Senécal, F. Morin, and J.-L. Gauvain. Neural probabilistic language models. In *Innovations in Machine Learning*, pages 137–186. Springer, 2006. 2

[3] Y. Bengio, P. Simard, and P. Frasconi. Learning long-term dependencies with gradient descent is difficult. *Neural Networks, IEEE Transactions on*, 5(2):157–166, 1994. 1, 2, 3

[4] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE, 2009. 4, 5

[5] D. Elliott and F. Keller. Comparing automatic evaluation measures for image description. 2014. 7

[6] J. L. Elman. Finding structure in time. *Cognitive science*, 14(2):179–211, 1990. 1

[7] 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. Springer, 2010. 2

[8] 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, 5, 6, 8

[9] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. 2014. 2, 5

[10] Y. Gong, L. Wang, M. Hodosh, J. Hockenmaier, and S. Lazebnik. Improving image-sentence embeddings using large weakly annotated photo collections. In *ECCV*, pages 529–545, 2014. 2

[11] A. Gupta, Y. Verma, and C. Jawahar. Choosing linguistics over vision to describe images. In *AAAI*, 2012. 2

[12] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 1, 2

[13] M. Hodosh, P. Young, and J. Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. *J. Artif. Intell. Res.(JAIR)*, 47:853–899, 2013. 1, 2, 6, 8

[14] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014. 4, 5

[15] M. A. Just, S. D. Newman, T. A. Keller, A. McEleney, and P. A. Carpenter. Imagery in sentence comprehension: an fmri study. *Neuroimage*, 21(1):112–124, 2004. 1

[16] A. Karpathy, A. Joulin, and L. Fei-Fei. Deep fragment embeddings for bidirectional image sentence mapping. *arXiv preprint arXiv:1406.5679*, 2014. 1, 2, 5, 6, 8

[17] R. Kiros, R. Salakhutdinov, and R. Zemel. Multimodal neural language models. In *ICML*, 2014. 2

[18] 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. 2, 4

[19] G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg. Baby talk: Understanding and generating simple image descriptions. In *CVPR*, pages 1601–1608. IEEE, 2011. 2, 5, 6, 8

[20] P. Kuznetsova, V. Ordonez, A. C. Berg, T. L. Berg, and Y. Choi. Collective generation of natural image descriptions. 2012. 2

[21] L. R. Lieberman and J. T. Culpepper. Words versus objects: Comparison of free verbal recall. *Psychological Reports*, 17(3):983–988, 1965. 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 *ECCV*, 2014. 1, 4

[23] J. Mao, W. Xu, Y. Yang, J. Wang, and A. L. Yuille. Explain images with multimodal recurrent neural networks. *arXiv preprint arXiv:1410.1090*, 2014. 2, 3, 6, 7, 8

[24] T. Mikolov. Recurrent neural network based language model. 1, 2, 3, 4, 5

[25] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *International Conference on Learning Representations: Workshops Track*, 2013. 3

[26] T. Mikolov, A. Deoras, D. Povey, L. Burget, and J. Cernocky. Strategies for training large scale neural network language models. In *Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on*, pages 196–201. IEEE, 2011. 4

[27] T. Mikolov and G. Zweig. Context dependent recurrent neural network language model. In *SLT*, pages 234–239, 2012. 1, 2, 3, 5

[28] M. Mitchell, X. Han, J. Dodge, A. Mensch, A. Goyal, A. Berg, K. Yamaguchi, T. Berg, K. Stratos, and H. Daumé III. Midge: Generating image descriptions from

computer vision detections. In *EACL*, pages 747–756. Association for Computational Linguistics, 2012. 2, 5, 6, 8

[29] A. Paivio, T. B. Rogers, and P. C. Smythe. Why are pictures easier to recall than words? *Psychonomic Science*, 11(4):137–138, 1968. 1

[30] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics, 2002. 1, 5

[31] C. Rashtchian, P. Young, M. Hodosh, and J. Hockenmaier. Collecting image annotations using Amazon's mechanical turk. In *NAACL HLT Workshop Creating Speech and Language Data with Amazon's Mechanical Turk*, 2010. 1, 4

[32] R. Socher, Q. Le, C. Manning, and A. Ng. Grounded compositional semantics for finding and describing images with sentences. In *NIPS Deep Learning Workshop*, 2013. 1, 2, 5, 6, 8

[33] I. Sutskever, J. Martens, and G. E. Hinton. Generating text with recurrent neural networks. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 1017–1024, 2011. 2

[34] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096–1103. ACM, 2008. 3

[35] R. J. Williams and D. Zipser. Experimental analysis of the real-time recurrent learning algorithm. *Connection Science*, 1(1):87–111, 1989. 4

[36] Y. Yang, C. L. Teo, H. Daumé III, and Y. Aloimonos. Corpus-guided sentence generation of natural images. In *EMNLP*, 2011. 2

[37] B. Z. Yao, X. Yang, L. Lin, M. W. Lee, and S.-C. Zhu. I2T: Image parsing to text description. *Proceedings of the IEEE*, 98(8):1485–1508, 2010. 2

[38] C. L. Zitnick and D. Parikh. Bringing semantics into focus using visual abstraction. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 3009–3016. IEEE, 2013. 2