# Multimodal Neural Language Models

**Ryan Kiros**
University of Toronto
rkiros@cs.toronto.edu

**Richard S. Zemel**
University of Toronto
zemel@cs.toronto.edu

**Ruslan Salakhutdinov**
University of Toronto
rsalakhu@cs.toronto.edu

## Abstract

We introduce two multimodal neural language models: models of natural language that can be conditioned on other modalities. A multimodal neural language model can be used to retrieve images given complex description queries, retrieve phrase descriptions given image queries, as well as generate text conditioned on images. We show that in the case of image-text modelling we can jointly learn word representations and image features by training our models together with a convolutional network. Unlike most existing methods, our approach can generate sentence descriptions for images without the use of templates, structured prediction, and/or syntactic trees. While we focus on image-text modelling, our algorithms can be easily applied to other modalities such as audio.

## 1 Introduction

Descriptive language is almost never isolated from other modalities. Advertisements come with images of the product that is being sold, social media profiles contain both descriptions and images of the user while multimedia websites that play audio and video have associated descriptions and opinions of the content. Consider the task of creating an advertisement to sell an item. An algorithm that can model both text descriptions and pictures of the item would allow a user to (a): search for pictures given a text description of the desired content (b): find similar item descriptions given uploaded images and (c): automatically generate text to describe the item given pictures. In a similar fashion, a user could write a description of a style of song and retrieve songs to match the query description or given a song, generate text to describe its style. What all these tasks have in common is the need to go beyond bag-of-word representations of text to model complex descriptions with an associated modality.

In this paper we introduce multimodal neural language models, models of natural language that can be conditioned on other modalities. A multimodal neural language model can account for all the previously described modelling challenges. Unlike most previous approaches to generating image descriptions, our model makes no use of templates, structured models, or syntactic trees. Instead, it relies on word representations learned from millions of words and conditioning the model on high-level image features learned from deep neural networks. We introduce two methods based on the log-bilinear model of [1]: the modality-biased log-bilinear model and the factored 3-way log-bilinear model. We then show how to learn word representations and image features together by jointly training our language models with a convolutional network. Experimentation is performed on two datasets with image-text descriptions: IAPR TC-12 and the Attributes Discovery dataset. We further illustrate the capabilities of our models through quantitative retrieval evaluation and visualizations of our results.

## 2 Related Work

Our related work can largely be separated into three groups: neural language models, image content description and multimodal representation learning. We describe each of these areas separately as well as indicate the relationships within these research areas.

**Neural Language Models:** A neural language model improves on $n$-gram language models by reducing the curse of dimensionality through the use of distributed word representations. Each word in the vocabulary is represented as a real-valued feature vector such that the cosine of the angles between these vectors is high for semantically similar words. Several models have been proposed based on feed-forward networks [2], log-bilinear models [1], skip-gram models [3] and recurrent neural networks [4, 5]. The downside to these types of models is that naively, they often require training times that are linear in the vocabulary size. To speed up training [6] explored the use of noise-contrastive estimation while [7] inferred trees over words.

**Image Description Generation:** A growing body of research has been proposed on how to generate realistic text descriptions given an image. [8] consider learning an intermediate meaning space to project image and sentence features allowing them to retrieve text from images and vice versa. [9] construct a CRF using unary potentials from objects, attributes and prepositions and high-order potentials from text corpora, using an n-gram model for decoding and templates for constraints. To allow for more descriptive and poetic generation, [10] propose the use of syntactic trees constructed from 700,000 Flickr images and text descriptions. For large scale description generation, [11] showed that non-parametric approaches are effective on a dataset of one million image-text captions. We note that unlike most existing work, our generated text comes directly from language model samples without any additional templates, structure, or constraints.

**Multimodal Representation Learning:** Deep learning methods have been successfully used to learn representations from multiple modalities. [12] proposed using deep autoencoders to learn features from audio and video, while [13] introduced the multimodal deep Boltzmann machine as a joint model of images and text. Unlike [13], our proposed models are conditional and go beyond bag-of-word features. More recently, [14] and [15] propose methods for mapping images into a text representation space learned from a language model that incorporates global context [16] or a skip-gram model [3], respectively . This allowed [14, 15] to perform zero-shot learning, generalizing to classes the model has never seen before. Similar to our work, [15] combine convolutional networks with a language model but our work instead focuses on text generation and retrieval as opposed to object classification.

The remainder of the paper is structured as follows. We first review the log-bilinear model of [1] as it forms the foundation for our work. We then introduce our two proposed models as well as how to perform joint image-text feature learning. Finally, we describe our experiments and results.

## 3 The Log-Bilinear Language Model (LBL)

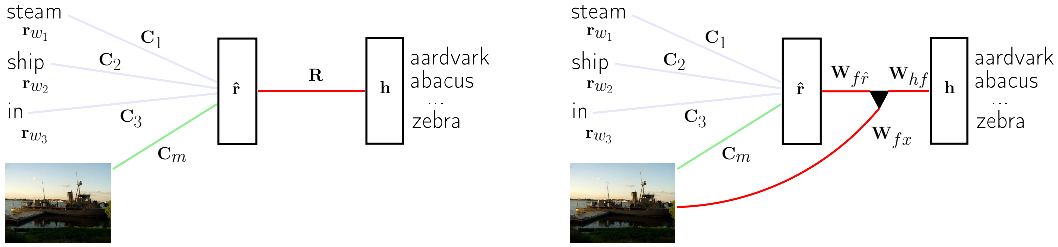
The log-bilinear language model (LBL) [1] is a deterministic model that may be viewed as a feed-forward neural network with a single linear hidden layer. As a neural language model, the LBL operates on word representation vectors. Each word $w$ in the vocabulary is represented as a $d$-dimension real-valued vector $\mathbf{r}_w \in \mathbb{R}^d$. Let $\mathbf{R}$ denote the $k \times d$ matrix of word representation vectors where $k$ is the vocabulary size. Let $(w_1, \ldots w_{n-1})$ be a tuple of $n-1$ words where $n-1$ is the context size. The LBL model makes a linear prediction of the next word representation as

$$\hat{\mathbf{r}} = \sum_{i=1}^{n-1} \mathbf{C}_i \mathbf{r}_{w_i} \tag{1}$$

where $\mathbf{C}_i, i = 1, \ldots, n-1$ are $d \times d$ context parameter matrices. Thus, $\hat{\mathbf{r}}$ is the predicted representation of $\mathbf{r}_{w_n}$. The conditional probability $P(w_n = w | w_{1:n-1})$ of $w_n$ given $w_1, \ldots, w_{n-1}$ is

$$P(w_n = w | w_{1:n-1}) = \frac{\exp(\hat{\mathbf{r}}^T \mathbf{r}_w + b_w)}{\sum_j \exp(\hat{\mathbf{r}}^T \mathbf{r}_j + b_j)} \tag{2}$$

where $\mathbf{b} \in \mathbb{R}^k$ is a bias vector with a word-specific bias $b_w$. Eq. 2 may be seen as scoring the predicted representation $\hat{\mathbf{r}}$ of $w_n$ against the actual representation $\mathbf{r}_w$ through an inner product, followed by normalization based on the inner products amongst all other word representations in the vocabulary. In the context of a feed-forward neural network, the weights between the output layer and linear hidden layer is the word representation matrix $\mathbf{R}$ where the output layer uses a softmax activation. Learning can be done with standard backpropagation.



(a) Modality-Biased Log-Bilinear Model (MLBL-B)

(b) Factored 3-way Log-Bilinear Model (MLBL-F)

Figure 1: Our proposed models. Left: The predicted next word representation $\hat{\mathbf{r}}$ is a linear prediction of word features $\mathbf{r}_{w_1}, \mathbf{r}_{w_2}, \mathbf{r}_{w_3}$ (blue connections) biased by image features $\mathbf{x}$. Right: The word representation matrix $\mathbf{R}$ is replaced by a factored tensor for which the hidden-to-output connections are gated by $\mathbf{x}$.

## 4 Multimodal Log-Bilinear Models

Suppose that along with each training tuple of words $(w_1, \ldots w_n)$ there is an associated vector $\mathbf{x} \in \mathbb{R}^m$ corresponding to the feature representation of the modality to be conditioned on, such as an image. Assume for now that these features are computed in advance. In section 5 we show how to jointly learn both text and image features.

### 4.1 Modality-Biased Log-Bilinear Model (MLBL-B)

Our first proposed model is the modality-biased log-bilinear model (MLBL-B) which is a straightforward extension of the LBL model. The MLBL-B model adds an additive bias to the next predicted word representation $\hat{\mathbf{r}}$ which is computed as

$$\hat{\mathbf{r}} = \left( \sum_{i=1}^{n-1} \mathbf{C}_i \mathbf{r}_{w_i} \right) + \mathbf{C}_m \mathbf{x} \tag{3}$$

where $\mathbf{C}_m$ is a $k \times m$ context matrix. Given the predicted next word representation $\hat{\mathbf{r}}$, computing the conditional probability $P(w_n = w | w_{1:n-1}, \mathbf{x})$ of $w_n$ given $w_1, \ldots, w_{n-1}$ and $\mathbf{x}$ is

$$P(w_n = w | w_{1:n-1}, \mathbf{x}) = \frac{\exp(\hat{\mathbf{r}}^T \mathbf{r}_w + b_w)}{\sum_j \exp(\hat{\mathbf{r}}^T \mathbf{r}_j + b_j)} \tag{4}$$

which remains unchanged from the LBL model. The MLBL-B can be viewed as a feedforward network by taking the LBL network and adding an additional context channel based on the modality $\mathbf{x}$, as shown in Fig. 1a. Learning in this model involves a straightforward application of backpropagation as in the LBL model.

### 4.2 The Factored 3-way Log-Bilinear Model (MLBL-F)

A more powerful model to incorporate modality conditioning is to gate the word representation matrix $\mathbf{R}$ by the features $\mathbf{x}$. By doing this, $\mathbf{R}$ becomes a tensor for which each feature $x$ can specify its own hidden to output weight matrix. More specifically, let $\mathbf{R}^{(1)}, \ldots, \mathbf{R}^{(m)}$ be $k \times d$ matrices specified by feature components $1, \ldots, m$ of $\mathbf{x}$. The hidden to output weights corresponding to $\mathbf{x}$ are computed as

$$\mathbf{R}_x = \sum_{i=1}^{m} x^{(i)} \mathbf{R}^{(i)} \tag{5}$$

where $\mathbf{R}_x$ denotes the word representations with respect to $\mathbf{x}$. The motivation for using a modality specific word representation is as follows. Suppose $\mathbf{x}$ is an image containing a picture of a cat, with context words $(\text{there}, \text{is}, \text{a})$. A language model that is trained without knowledge of image features would score the predicted next word representation $\hat{\mathbf{r}}$ high with words such as dog, building or car. If each image has a corresponding word representation matrix $\mathbf{R}_x$, the representations for attributes that are not present in the image would be modified such that the inner product of $\hat{\mathbf{r}}$ with the representation of cat would score higher than the inner product of $\hat{\mathbf{r}}$ with the representations of dog, building or car.

As is, the tensor $\mathbf{R}$ requires $k \times d \times m$ parameters which makes using a general 3-way tensor impractical even for modest vocabulary sizes. A common solution to this approach [17, 18] is to factor $\mathbf{R}$ into 3 matrices $\mathbf{W}_{f\hat{r}}$, $\mathbf{W}_{fx}$ and $\mathbf{W}_{hf}$ such that

$$\mathbf{R}_x^T = \mathbf{W}_{hf} \odot \delta(\mathbf{W}_{fx}\mathbf{x}) \odot \mathbf{W}_{f\hat{r}} \tag{6}$$

where $\delta(\cdot)$ denotes the matrix with its argument on the diagonal and $\odot$ is the Hadamard product. These matrices are parametrized by $f$, the number of factors, as shown in Fig. 1b.

Under this model, the predicted next word representation $\hat{\mathbf{r}}$ is

$$\hat{\mathbf{r}} = \left( \sum_{i=1}^{n-1} \mathbf{C}_i (\mathbf{W}_{hf}\mathbf{W}_{f\hat{r}})_{w_i}^T \right) + \mathbf{C}_m \mathbf{x} \tag{7}$$

where $(\mathbf{W}_{hf}\mathbf{W}_{f\hat{r}})_{w_i}^T$ denotes the column of $(\mathbf{W}_{hf}\mathbf{W}_{f\hat{r}})^T$ for the word representation of $w_i$. Given a predicted next word representation $\hat{\mathbf{r}}$, the factor outputs are

$$\mathbf{f} = (\mathbf{W}_{f\hat{r}}\hat{\mathbf{r}}) \odot (\mathbf{W}_{fx}\mathbf{x}) \tag{8}$$

with the conditional probability $P(w_n = w | w_{1:n-1}, \mathbf{x})$ of $w_n$ given $w_1, \ldots, w_{n-1}$ and $\mathbf{x}$ given by

$$P(w_n = w | w_{1:n-1}, \mathbf{x}) = \frac{\exp(\mathbf{W}_{hf}^{(w)}\mathbf{f} + b_w)}{\sum_j \exp(\mathbf{W}_{hf}^{(j)}\mathbf{f} + b_j)} \tag{9}$$

where $\mathbf{W}_{hf}^{(w)}$ denotes the row of $\mathbf{W}_{hf}$ corresponding to word $w$. We call this the MLBL-F model. As with the LBL and MLBL-B models, training can be achieved using a straightforward application of backpropagation. Unlike the other models, extra care needs to be taken when adjusting the learning rates for the matrices of the factored tensor.

It is sensible that pre-computed word embeddings could be used as a starting point to training, as opposed to random initialization of the word representations. Indeed, all of our experiments use the embeddings of [19] for initialization. In the case of the LBL and MLBL-B models, each pre-trained word embedding can be used to initialize the rows of $\mathbf{R}$. In the case of the MLBL-F model where $\mathbf{R}$ is a factored tensor, let $\mathbf{E}$ denote the $k \times d$ matrix of pre-trained embeddings and observe that $\mathbf{E} = (\mathbf{W}_{hf}\mathbf{W}_{f\hat{r}})^T$. Thus to initialize the MLBL-F model with pre-trained embeddings, we simply apply an SVD to $\mathbf{E}$.

## 5 Joint Image-Text Feature Learning

Up until now we have not made any assumptions on the type of modality being used for the feature representation $\mathbf{x}$. In this section, we consider the case where the conditioned modality consists of images and show how to jointly learn image and word features along with the model parameters.

One way of incorporating image representation learning is to use a convolutional network for which the outputs are used either as an additive bias or for gating. Gradients from the loss could then be backpropagated from the language model through the convolutional network to update filter weights. Unfortunately, this architecture posses some issues. Since each training tuple of words comes with an associative image, then the number of training elements becomes large even with a modest size training set. For example, if the training set consisted of 10,000 images and each image had a text description of 20 words, then the number of training instances for the model becomes roughly 200,000. For large image databases this could quickly scale to millions of training instances.

To speed up computation, we follow [20, 21] and choose to learn our convolutional networks on small feature maps learned using $k$-means as opposed to the original images. We follow the pipeline of [22]. Given training images, $r \times r$ patches are randomly extracted, contrast normalized and whitened. These are used for training a dictionary with spherical $k$-means. These filters are convolved with the image and a soft activation encoding is applied. If the image is of dimensions $n_V \times n_H \times 3$ and $k_f$ filters are learned, the resulting feature maps are of size $(n_V - r + 1) \times (n_H - r + 1) \times k_f$. Each slice of this region is then split into a $G \times G$ grid for which features within each region are max-pooled. This results in an output of size $G \times G \times k_f$. It is these outputs that are used as inputs to the convolutional network. For all of our experiments, we use $G = 9$ and $k_f = 128$.

Each $9 \times 9 \times 128$ input is convolved with 64 $3 \times 3$ filters resulting in feature maps of size $7 \times 7 \times 64$. Rectified linear units (ReLUs) are used for activation followed by a response normalization layer [23]. The response-normalized feature maps are then max-pooled with a pooling window of $3 \times 3$ and a stride of 2, resulting in outputs of size $3 \times 3 \times 64$. One fully-connected layer with ReLU activation is added. It is the feature responses at this layer that are used either for additive biasing or gating in the MLBL-B and MLBL-F models, respectively.

## 6 Retrieval and Generation

In this section we describe how our models can perform generation and retrieval. The standard approach to evaluating language models is through perplexity

$$\log_2 \mathcal{C}(w_{1:n}|\mathbf{x}) = -\frac{1}{N} \sum_{w_{1:n}} \log_2 P(w_n = w | w_{1:n-1}, \mathbf{x}) \tag{10}$$

where $w_{1:n-1}$ runs through each subsequence of length $n-1$ and $N$ is the length of the sequence. Here we use perplexity not only as a measure of performance but also as a link between both text and the additional modality.

First, consider the task of retrieving training images from a test description query $w_{1:N}$. For each image $\mathbf{x}$ in the training set, we compute $\mathcal{C}(w_{1:N}|\mathbf{x})$ and return the images for which $\mathcal{C}(w_{1:N}|\mathbf{x})$ is lowest. Intuitively, images when conditioned on by the model that achieve low perplexity are those that are a good match to the query description.

The task of retrieving text from an image query is trickier for the following reasons. It is likely that there are many 'easy' descriptions for which the language model will assign low perplexity to independent of the query image being conditioned on. Thus, instead of retrieving text from the training set for which $\mathcal{C}(w_{1:N}|\mathbf{x})$ is lowest conditioned on the query image $\mathbf{x}$, we instead look at the ratio $\mathcal{C}(w_{1:N}|\mathbf{x})/\mathcal{C}(w_{1:N}|\tilde{\mathbf{x}})$ where $\tilde{x}$ denotes the mean image in the training set (computed in feature space). Thus, if $w_{1:N}$ is a good explanation of $\mathbf{x}$, then $\mathcal{C}(w_{1:N}|\mathbf{x}) < \mathcal{C}(w_{1:N}|\tilde{\mathbf{x}})$ and we can simply retrieve the text for which this ratio is smallest.

While this leads to better search results, it is conceivable that using the image itself as a query for other images and returning their corresponding descriptions may in itself work well as a query strategy. For example, an image taken at night would ideally return a description describing this, which would be more likely to occur if we first retrieved nearby images which were also taken at night. We found the most effective way of performing description retrieval is as follows: first retrieve the top $k_r$ training images as a shortlist based on the Euclidean distance of $\mathbf{x}$ and images in the training set (in feature space). Then retrieve the descriptions for which $\mathcal{C}(w_{1:N}|\mathbf{x})/\mathcal{C}(w_{1:N}|\tilde{\mathbf{x}})$ is smallest for each description $w_{1:N}$ in the shortlist. We found that combining these two strategies is more effective than using either alone. In the case when a convolutional network is used, we first map the images through the convolutional network and use the output representations for computing distances.

Finally, we generate text given an image as follows: Suppose we are given an initialization $w_{1:n_i}$. We compute $P(w_n = w|w_{1:n_i}, \mathbf{x})$ and obtain a sample $\tilde{w}$ from this distribution, appending $\tilde{w}$ to our initialization. This procedure is then repeated for as long as desired.

## 7 Experiments

We perform experimental evaluation of our proposed models on two publicly available datasets:

**IAPR TC-12** This data set consists of 20,000 images across various domains, such as landscapes, portraits, indoor and sports scenes. Accompanied by each image is a text description of one to three sentences describing the content of the image. The dataset was initially released for cross-lingual retrieval [24] but has since been used extensively for other tasks such as image annotation. We used the publicly available train/test split for our experiments.

**Attribute Discovery** This dataset contains roughly 40,000 images related to products such as bags, clothing and shoes as well as subcategories of each product, such as high-heels and sneakers. Each image is accompanied by a web-retrieved text description which often reads as an advertisement for the product. Unlike the IAPR dataset, the text descriptions are not necessarily guaranteed to be descriptive of the image and often contains noisy, unrelated text. This dataset was proposed as a means of discovering visual attributes from noisy text [25]. We used a random train/test split for our experiments which will be made publicly available.

We chose to evaluate on these two datasets since they complement each other. The IAPR dataset contains a wide variety of image scenes but the associated text is clean, descriptive and follows a loose style that is consistent across images. On the other hand, the Attribute Discovery dataset contains images which are easy to learn from: mostly centered on a white background but are associated with noisy descriptions which are arguably more realistic to image data on the web.

### 7.1 Details of Experiments

We perform four experiments, three of which are quantitative and one of which is qualitative:

**Bleu Evaluation** Our main evaluation criteria is based on Bleu [26]. Bleu was designed for automated evaluation of statistical machine translation and can be used in our setting to measure similarity of descriptions. Previous work on generating text descriptions for images use Bleu as a means of evaluation, where the generated description is used as a candidate for the gold standard reference generation. Given the diversity of possible image descriptions, Bleu may penalize candidates which are arguably descriptive of image content as noted by [9], though Bleu remains the standard evaluation criteria for such models. Given a model, we generate a candidate description as described in section 6, generating as many words as there are in the reference description and compute the Bleu score of the candidate with the reference. This is repeated over all test points ten times, in order to account for the variability in the generated descriptions. For baselines, we also compare against the log-bilinear model as well as image-conditioned models conditioned on random images as a control. This allows us to obtain further evidence of the relevance of generated text. Finally, we compare against the models of [27] and [28] who report Bleu scores for their models on the IAPR dataset. [1]

**Perplexity Evaluation** Each of our proposed models are trained on both datasets and the perplexity of the language models are evaluated. As baselines, we also include the basic log-bilinear model as well as two n-gram models. To evaluate the effectiveness of using pre-trained word embeddings, we also train a log-bilinear model where the word representations are randomly initialized. We hypothesize that image-conditioned models should result in lower perplexity than models which are only trained on text without knowledge of their associate images.

**Retrieval Evaluation** We quantitatively evaluate the performance of our model for doing retrieval. First consider the task of retrieving images from description queries. Given a test description, we compute the model perplexity conditioned on each test image and rank each image accordingly. Let $k_r$ denote the number of retrieved images. We define a description to be correctly matched if the matching image to the description query is ranked in the top $k_r$ images sorted by model perplexity. Retrieving descriptions from image queries is performed equivalently. Since our models use a shortlist (see section 6) of nearest images for retrieving descriptions, we restrict our search to images within the shortlist, for which the matching description is guaranteed to be in.

For additional comparison, we include a bag-of-words baseline to determine whether a language model (and word ordering) is necessary for image-description retrieval tasks. This model works as follows: given image features (either the features of [29] or features learned through $k$-means), we learn a projection onto independent logistic units, one for each word in the description. The score of a description is then the sum of the probabilities of each word in the description, normalized by the number of words. For retrieving images, we project each image and rank those which result in the highest description score. For retrieving descriptions, we return those which result in the highest score given the word probabilities computed from the image. Since we use a shortlist for our models when performing description retrieval, we also use an equivalent shortlist for the baseline model to allow for fair comparison. Training the baseline model was done using SGD with minibatch sizes of 50, with a validation set used to tune the amount of weight decay.

**Qualitative Results** Finally, we qualitatively evaluate image to text retrieval, text to image retrieval as well as samples from our models. Several additional qualitative results are displayed in the appendix.

### 7.2 Details of Training

Each of our language models were trained using the same hyperparameters: all context matrices used a weight decay of $1.0 \times 10^{-4}$ while word representations used a weight decay of $1.0 \times 10^{-5}$. All

[1] We note that an exact comparison cannot be made with these methods since [28] assume tags are given as input along with images and both methods apply 10-fold CV. The use of tags can substantially boost the relevance of generated descriptions. None the less, these methods provide us context for the results of our models.

other weight matrices, including the convolutional network filters use a weight decay of $1.0 \times 10^{-4}$. We used batch sizes of 20 and an initial learning rate of 0.2 which was exponentially decreased at each epoch by a factor of 0.998. Gated methods used an initial learning rate of 0.02. Initial momentum was set to 0.5 and is increased linearly to 0.9 over 20 epochs. The word representation matrices were initialized to the pre-trained embeddings of [19] with all other weights randomly initialized from a zero mean Gaussian with a standard deviation of 0.01. We used a context size of 5 for each of our models. Non-convolutional methods used the fixed image features of [29] which are 256 dimensional. For fair comparison, our convolutional models also used a 256-dimensional output layer. Perplexity was computed starting with word $C + 1$ for all methods where $C$ is the largest context size used in comparison (5 in our experiments). Perplexity was not evaluated on descriptions shorter than $C + 3$ words.

For each of our experiments, we split the training set into 80% training and 20% validation. Each model was trained while monitoring the perplexity on the validation set. Once the perplexity no longer improved for 5 epochs, the objective value on the training set was recorded. The training and validation sets were then fused and training continued until the objective value on the validation batch matched the recorded training objective. At this point, training stopped and evaluation was performed on the test set.

### 7.3 Generation and Perplexity Results

Table 1 shows results on the IAPR and Attributes dataset, respectively. On both datasets, each of our multimodal models outperforms both the log-bilinear and n-gram models on Bleu scores. What is perhaps most surprising is that simple language models independent of images can also achieve non-trivial Bleu scores. For further comparison, we also computed Bleu scores on the convolutional MLBL-B model when random images are used for conditioning. This gives us strong evidence that the gains in Bleu scores are directly from capturing and associating word representations from image content. [2]

One observation from our results is that perplexity does not appear to be correlated with Bleu scores. On the IAPR dataset, the best perplexity is obtained using the MLBL-B model with fixed features, even though the best Bleu scores are obtained with a convolutional model. Similarly, both Back-off GT3 and LBL have the lowest perplexities on the Attributes dataset but are worse with respect to Bleu. Using more than 3-grams did not improve results on either dataset. The combination of perplexity and Bleu may better be seen as measuring readability and relevance, respectively. For additional comparison, we also ran an experiment training LBL on both datasets using random word initialization, achieving perplexity scores of 23.4 and 109.6. This indicates the benefit of initialization from pre-trained word representations. Perhaps unsurprisingly, perplexity is much worse on the convolutional MLBL-B model when random images are used for conditioning.

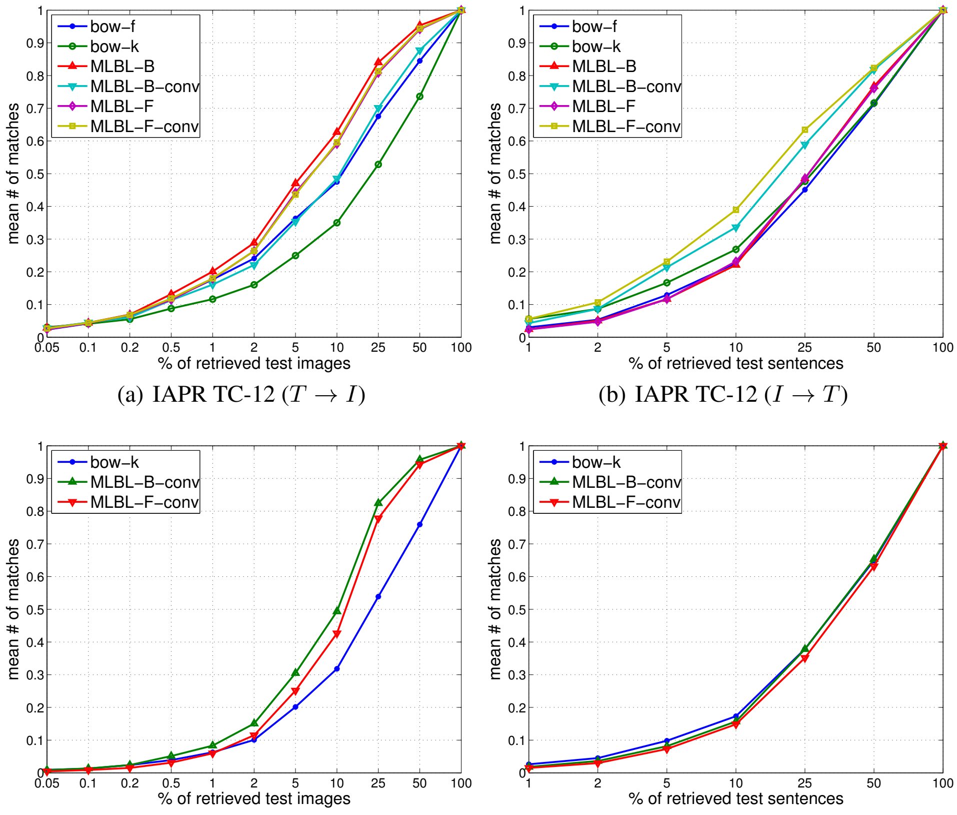
Table 1: Results on IAPR (top) and Attributes (bottom). PPL refers to perplexity while B-n indicates Bleu scored with $n$-grams. Back-off GTn refers to $n$-grams with Good-Turing discounting. Models which use a convolutional network are indicated by -conv, while -conv-R indicates using random images for conditioning. The MLBL-B and MLBL-F models use the fixed publicly available features from [29].

| Model type | PPL. | 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 | 20.1 | 0.327 | 0.144 | 0.068 |
| MLBL-B-conv-R | 28.7 | 0.325 | 0.143 | 0.069 |
| MLBL-B | **18.0** | **0.349** | 0.161 | 0.079 |
| MLBL-F | 20.3 | 0.348 | 0.165 | **0.085** |
| MLBL-B-conv | 20.6 | **0.349** | 0.165 | **0.085** |
| MLBL-F-conv | 21.7 | 0.341 | 0.156 | 0.073 |
| Gupta et al. 1 | - | 0.15 | 0.06 | 0.01 |
| Gupta & Mannem 1 | - | 0.33 | **0.18** | 0.07 |
| Back-off GT2 | 117.7 | 0.163 | 0.033 | 0.009 |
| Back-off GT3 | **93.4** | 0.166 | 0.032 | 0.011 |
| LBL | 97.6 | 0.161 | 0.031 | 0.009 |
| MLBL-B-conv-R | 154.4 | 0.166 | 0.035 | 0.012 |
| MLBL-B-conv | 99.2 | **0.189** | **0.048** | **0.017** |
| MLBL-F-conv | 113.2 | 0.175 | 0.042 | 0.014 |

### 7.4 Retrieval Results

Figure 2 illustrates the results of our retrieval experiments. We observe that all models perform comparatively when retrieving from a small percentage of the test set, while our models are able to outperform the baseline when retrieving from a larger percentage on all experiments except for description retrieval on the Attributes dataset. Interestingly, both convolutional models outperform

[2] We also evaluted Bleu scores after stopword removal and observed the same trend in performance.



(a) IAPR TC-12 ($T \rightarrow I$)

(b) IAPR TC-12 ($I \rightarrow T$)

(c) Attributes ($T \rightarrow I$)

(d) Attributes ($I \rightarrow T$)

Figure 2: Retrieval results on IAPR TC-12 (top) and the Attributes dataset (bottom) when using descriptions to retrieve images ($T \rightarrow I$, left) and images to retrieve descriptions ($I \rightarrow T$, right). bow-f refers to the bag of words baseline applied to the features of [29] and bow-k equivalently for $k$-means feature maps (the inputs used to our convolutional models). Retrieving descriptions from image queries (right) use a shortlist size of 100. For the Attributes dataset, we used a random subset of 5000 test images.

all other approaches for description retrieval on IAPR TC-12, while the additive model performs the best when retrieving images. We hypothesize that due to the noisy text present in the Attributes dataset, using a language model for retrieving text would likely have little advantage over a bag of words model in this scenario. This may partially explain the outcome of figure 2(d).

### 7.5 Qualitative results

On the left of Figure 3, we illustrate sample results for retrieving text from images. Each test image is displayed along with the top query for which the perplexity was lowest on a shortlist of 15 images. In general, the model does a good job at retrieving text with general characteristics of a scene or retrieving the correct type of product on the Attributes Discovery dataset, being able to distinguish between different kinds of sub-products, such as shoes and boots. The most common mistakes that the model makes are retrieving text with extraneous descriptions that do not exist in the image, such as describing people that are not present. On the right of Figure 3 are sample text generation results. The model was initialized with 'in this picture there is' or 'this product contains a' and proceeded to generate 50 words conditioned on the image. We generated 5-10 examples and show the result we found to be either the most accurate or amusing. The model is often able to describe the general content of the image, even if it does not get specifics correct such as colors of clothing. This gives visual confirmation of the increased Bleu scores from our models.

We also illustrate results of retrieving images given a text query in figure 4. For each test description, we retrieved the top 4 training images for which the perplexity was lowest when conditioned on.

The bottom-right image depicts a case where the model completely misses the desired result. We observed that this occurs most often on shorter queries where single words, such as sunset and lake, indicate key visual concepts that the model is not able to pick up on. Contrast this to the shoe images on the top left, where the model can correctly identify the difference between a sneaker and a clog, even though these two words only appear once or twice in the description.

# 8 Conclusion

In this paper we proposed multimodal neural language models. We described two novel language models and showed in the context of image-text learning how to jointly learn word representations and image features. Interestingly, our models can obtain comparable Bleu scores to existing approaches for description generation simply from sampling from the model while improving over a strong bag-of-words baseline for description and image retrieval.

From our experiments, it is not immediately clear the advantages of our proposed models in comparison to each other and whether or not incorporating a convolutional net offers any significant advantage over a fixed image representation. We strongly suspect the success of the additive bias models are due in part to use of high-level image feature representations learned from multi-layer architectures. Recently, [30] showed that a convolutional network trained on ImageNet can be used as a general feature extractor by computing features from the top fully connected layers. We anticipate that using these fixed feature representations would be the most effective for our models.

This work takes a first step towards generating image descriptions with a multimodal language model. Ideally, we should be able to take into account interactions of objects within scenes. Consider two pictures: one with a cat sitting on a box and the other with a box on top of a cat. We would like to be able to extend our models to learn latent representations of these interactions and incorporate them into our models to obtain better fine-grained descriptions. We suspect that in these tasks, the theoretical strengths of the multiplicative models will become much more apparent.

### Acknowledgments

This research was supported by OGS and NSERC. We would also like to thank the anonymous reviewers for their valuable comments and suggestions.

## References

[1] Andriy Mnih and Geoffrey Hinton. Three new graphical models for statistical language modelling. In *Proceedings of the 24th international conference on Machine learning*, pages 641–648. ACM, 2007.

[2] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. A neural probabilistic language model. *J. Mach. Learn. Res.*, 3:1137–1155, March 2003.

[3] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[4] Tomas Mikolov, Martin Karafiát, Lukas Burget, Jan Cernockỳ, and Sanjeev Khudanpur. Recurrent neural network based language model. In *INTERSPEECH*, pages 1045–1048, 2010.

[5] Tomas Mikolov, Anoop Deoras, Daniel Povey, Lukas Burget, and Jan 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.

[6] Andriy Mnih and Yee Whye Teh. A fast and simple algorithm for training neural probabilistic language models. *arXiv preprint arXiv:1206.6426*, 2012.

[7] Andriy Mnih and Geoffrey E Hinton. A scalable hierarchical distributed language model. In *Advances in neural information processing systems*, pages 1081–1088, 2008.

[8] Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. Every picture tells a story: Generating sentences from images. In *Computer Vision–ECCV 2010*, pages 15–29. Springer, 2010.

[9] Girish Kulkarni, Visruth Premraj, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C Berg, and Tamara L Berg. Baby talk: Understanding and generating simple image descriptions. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1601–1608. IEEE, 2011.

[10] Margaret Mitchell, Xufeng Han, Jesse Dodge, Alyssa Mensch, Amit Goyal, Alex Berg, Kota Yamaguchi, Tamara Berg, Karl Stratos, and Hal Daumé III. Midge: Generating image descriptions from computer vision detections. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 747–756. Association for Computational Linguistics, 2012.

[11] Vicente Ordonez, Girish Kulkarni, and Tamara L Berg. Im2text: Describing images using 1 million captioned photographs. In *Advances in Neural Information Processing Systems*, pages 1143–1151, 2011.

[12] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Ng. Multimodal deep learning. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 689–696, 2011.

[13] Nitish Srivastava and Ruslan Salakhutdinov. Multimodal learning with deep boltzmann machines. In *Advances in Neural Information Processing Systems 25*, pages 2231–2239, 2012.

[14] Richard Socher, Milind Ganjoo, Hamsa Sridhar, Osbert Bastani, Christopher D Manning, and Andrew Y Ng. Zero-shot learning through cross-modal transfer. *arXiv preprint arXiv:1301.3666*, 2013.

[15] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeffrey Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. 2013.

[16] Eric H Huang, Richard Socher, Christopher D Manning, and Andrew Y Ng. Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 873–882. Association for Computational Linguistics, 2012.

[17] Roland Memisevic and Geoffrey Hinton. Unsupervised learning of image transformations. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE, 2007.

[18] Alex Krizhevsky, Geoffrey E Hinton, et al. Factored 3-way restricted boltzmann machines for modeling natural images. In *International Conference on Artificial Intelligence and Statistics*, pages 621–628, 2010.

[19] Ronan Collobert and Jason Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167. ACM, 2008.

[20] Tao Wang, David J Wu, Adam Coates, and Andrew Y Ng. End-to-end text recognition with convolutional neural networks. In *Pattern Recognition (ICPR), 2012 21st International Conference on*, pages 3304–3308. IEEE, 2012.

[21] Kevin Swersky, Jasper Snoek, and Ryan Adams. Multi-task bayesian optimization. In *Advances in Neural Information Processing Systems*, 2013.

[22] Adam Coates and Andrew Ng. The importance of encoding versus training with sparse coding and vector quantization. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 921–928, 2011.

[23] Alex Krizhevsky, Ilya Sutskever, and Geoff Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25*, pages 1106–1114, 2012.

[24] Michael Grubinger, Paul Clough, Henning Müller, and Thomas Deselaers. The iapr tc-12 benchmark: A new evaluation resource for visual information systems. In *International Workshop OntoImage*, pages 13–23, 2006.

[25] Tamara L Berg, Alexander C Berg, and Jonathan Shih. Automatic attribute discovery and characterization from noisy web data. In *Computer Vision–ECCV 2010*, pages 663–676. Springer, 2010.

[26] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing 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.

[27] Ankush Gupta, Yashaswi Verma, and CV Jawahar. Choosing linguistics over vision to describe images. In *AAAI*, 2012.

[28] Ankush Gupta and Prashanth Mannem. From image annotation to image description. In *Neural Information Processing*, pages 196–204. Springer, 2012.

[29] Ryan Kiros and Csaba Szepesvári. Deep representations and codes for image auto-annotation. In *Advances in Neural Information Processing Systems 25*, pages 917–925, 2012.

[30] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. *arXiv preprint arXiv:1310.1531*, 2013.

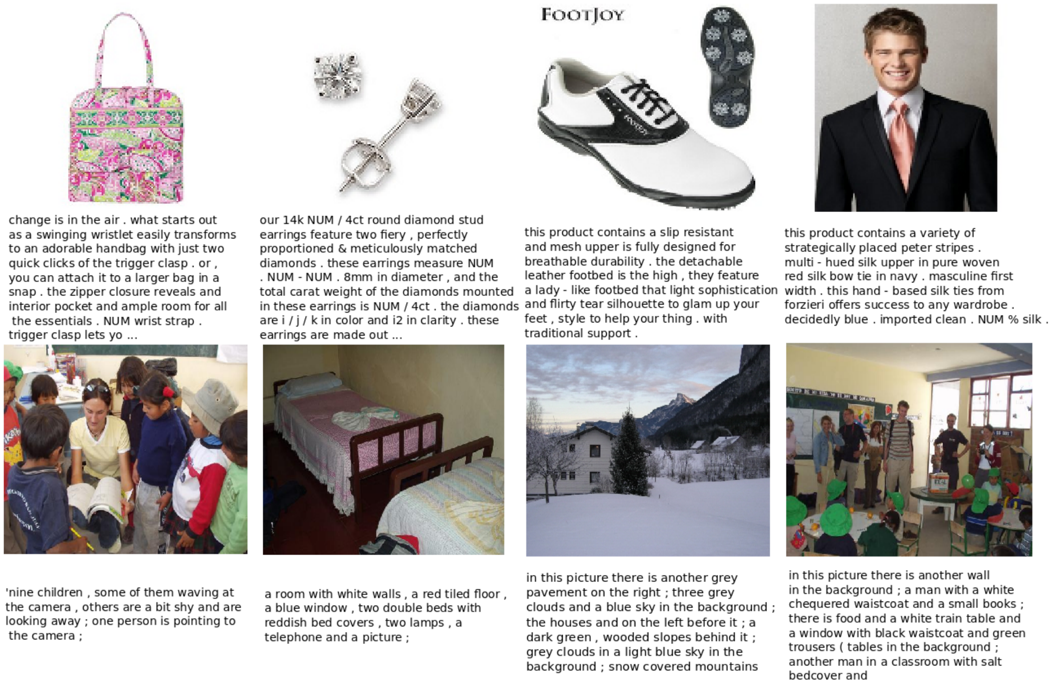

FOOTJOY
change is in the air . what starts out as a swinging wristlet easily transforms to an adorable handbag with just two quick clicks of the trigger clasp . or , you can attach it to a larger bag in a snap . the zipper closure reveals and interior pocket and ample room for all the essentials . NUM wrist strap . trigger clasp lets yo ...
our 14k NUM / 4ct round diamond stud earrings feature two fiery , perfectly proportioned & meticulously matched diamonds . these earrings measure NUM . NUM - NUM . 8mm in diameter , and the total carat weight of the diamonds mounted in these earrings is NUM / 4ct . the diamonds are i / j / k in color and i2 in clarity . these earrings are made out ...
this product contains a slip resistant and mesh upper is fully designed for breathable durability . the detachable leather footbed is the high , they feature a lady - like footbed that light sophistication and flirty tear silhouette to glam up your feet , style to help your thing . with traditional support .
this product contains a variety of strategically placed peter stripes . multi - hued silk upper in pure woven red silk bow tie in navy . masculine first width . this hand - based silk ties from forzieri offers success to any wardrobe . decidedly blue . imported clean . NUM % silk .
'nine children , some of them waving at the camera , others are a bit shy and are looking away ; one person is pointing to the camera ;
a room with white walls , a red tiled floor , a blue window , two double beds with reddish bed covers , two lamps , a telephone and a picture ;
in this picture there is another grey pavement on the right ; three grey clouds and a blue sky in the background ; the houses and on the left before it ; a dark green , wooded slopes behind it ; grey clouds in a light blue sky in the background ; snow covered mountains
in this picture there is another wall in the background ; a man with a white chequered waistcoat and a small books ; there is food and a white train table and a window with black waistcoat and green trousers ( tables in the background ; another man in a classroom with salt bedcover and

Figure 3: Left: Description retrieval given images, returning the top query. Right: Description generation. Descriptions were initialized with either 'in this picture there is' or 'this product contains a'.

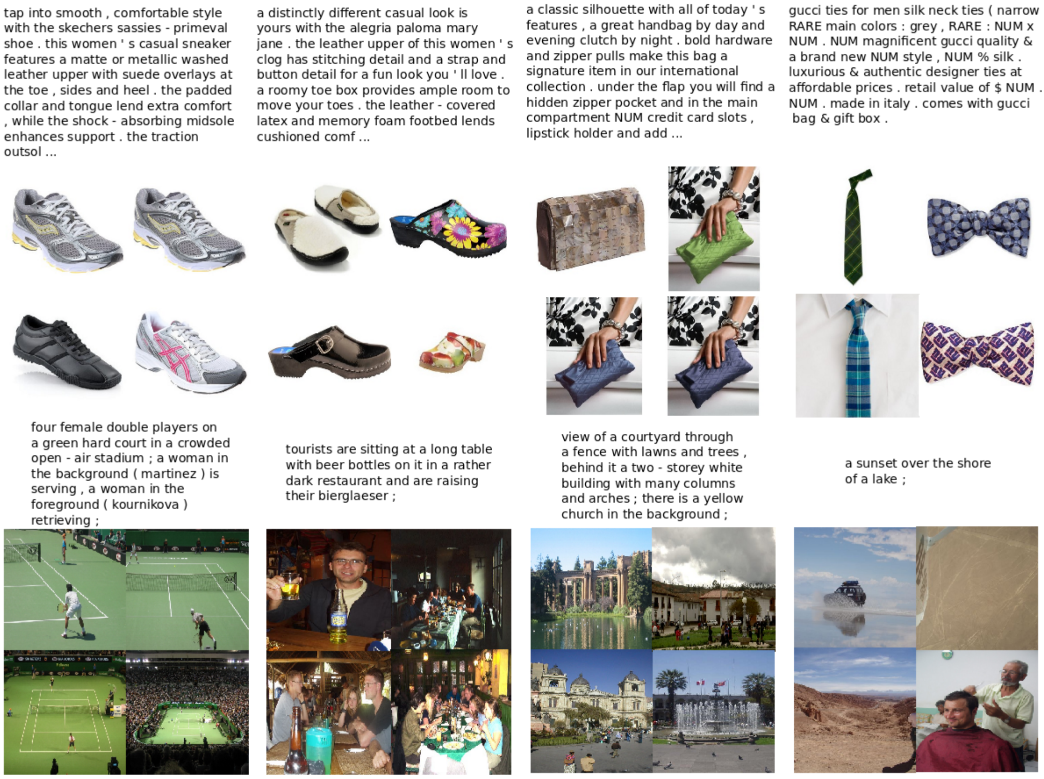
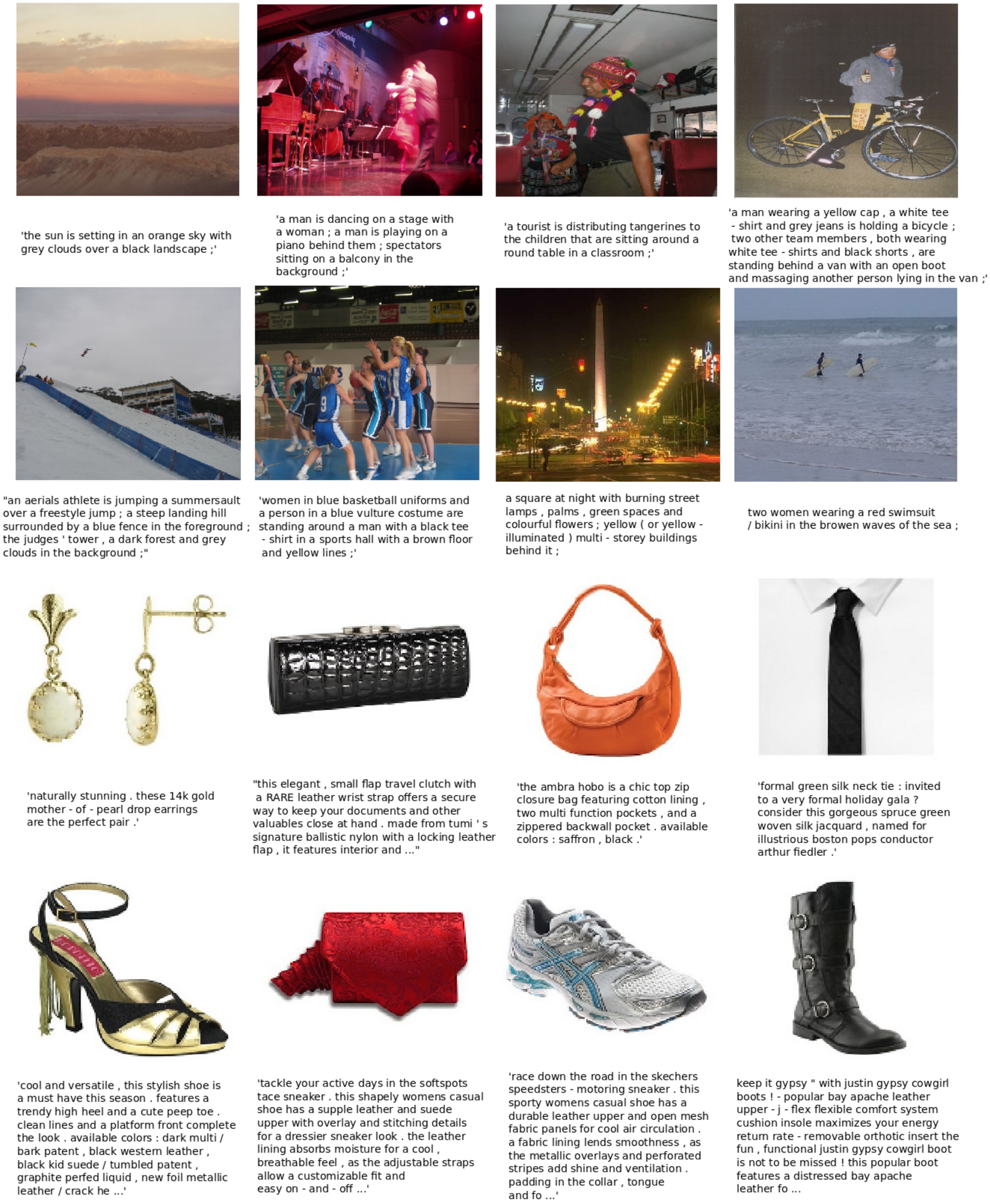

tap into smooth , comfortable style with the skechers sassies - primeval shoe . this women ' s casual sneaker features a matte or metallic washed leather upper with suede overlays at the toe , sides and heel . the padded collar and tongue lend extra comfort , while the shock - absorbing midsole enhances support . the traction outsol ...
a distinctly different casual look is yours with the alegria paloma mary jane . the leather upper of this women ' s clog has stitching detail and a strap and button detail for a fun look you ' ll love . a roomy toe box provides ample room to move your toes . the leather - covered latex and memory foam footbed lends cushioned comf ...
a classic silhouette with all of today ' s features , a great handbag by day and evening clutch by night . bold hardware and zipper pulls make this bag a signature item in our international collection . under the flap you will find a hidden zipper pocket and in the main compartment NUM credit card slots , lipstick holder and add ...
gucci ties for men silk neck ties ( narrow ) RARE main colors : grey , RARE : NUM x NUM . NUM magnificent gucci quality & a brand new NUM style , NUM % silk . luxurious & authentic designer ties at affordable prices . retail value of $ NUM . NUM . made in italy . comes with gucci bag & gift box .
four female double players on a green hard court in a crowded open - air stadium ; a woman in the background ( martinez ) is serving , a woman in the foreground ( kournikova ) retrieving ;
tourists are sitting at a long table with beer bottles on it in a rather dark restaurant and are raising their bierglaeser ;
view of a courtyard through a fence with lawns and trees , behind it a two - storey white building with many columns and arches ; there is a yellow church in the background ;
a sunset over the shore of a lake ;

Figure 4: Sample image retrieval given descriptions. The top 4 results are displayed.

## Additional qualitative results: description retrieval



'the sun is setting in an orange sky with grey clouds over a black landscape ;'

'a man is dancing on a stage with a woman ; a man is playing on a piano behind them ; spectators sitting on a balcony in the background ;'

'a tourist is distributing tangerines to the children that are sitting around a round table in a classroom ;'

'a man wearing a yellow cap , a white tee - shirt and grey jeans is holding a bicycle ; two other team members , both wearing white tee - shirts and black shorts , are standing behind a van with an open boot and massaging another person lying in the van ;'

"an aerials athlete is jumping a summersault over a freestyle jump ; a steep landing hill surrounded by a blue fence in the foreground ; the judges ' tower , a dark forest and grey clouds in the background ;"

'women in blue basketball uniforms and a person in a blue vulture costume are standing around a man with a black tee - shirt in a sports hall with a brown floor and yellow lines ;'

a square at night with burning street lamps , palms , green spaces and colourful flowers ; yellow ( or yellow - illuminated ) multi - storey buildings behind it ;

two women wearing a red swimsuit / bikini in the browen waves of the sea ;

'naturally stunning . these 14k gold mother - of - pearl drop earrings are the perfect pair .'

"this elegant , small flap travel clutch with a RARE leather wrist strap offers a secure way to keep your documents and other valuables close at hand . made from tumi ' s signature ballistic nylon with a locking leather flap , it features interior and ..."

'the ambra hobo is a chic top zip closure bag featuring cotton lining , two multi function pockets , and a zippered backwall pocket . available colors : saffron , black .'

'formal green silk neck tie : invited to a very formal holiday gala ? consider this gorgeous spruce green woven silk jacquard , named for illustrious boston pops conductor arthur fiedler .'

'cool and versatile , this stylish shoe is a must have this season . features a trendy high heel and a cute peep toe . clean lines and a platform front complete the look . available colors : dark multi / bark patent , black western leather , black kid suede / tumbled patent , graphite perfed liquid , new foil metallic leather / crack he ...'

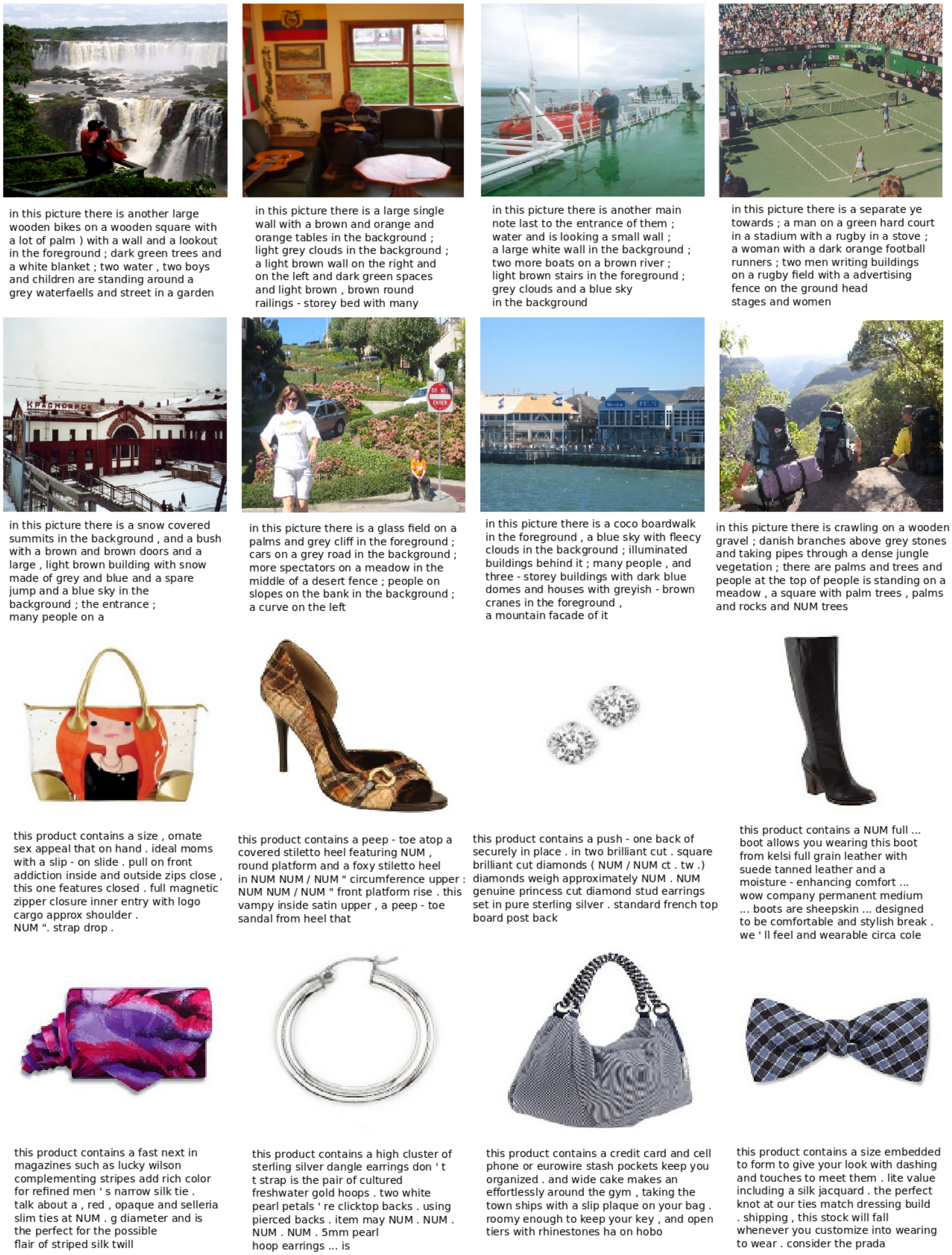
'tackle your active days in the softspots tace sneaker . this shapely womens casual shoe has a supple leather and suede upper with overlay and stitching details for a dressier sneaker look . the leather lining absorbs moisture for a cool , breathable feel , as the adjustable straps allow a customizable fit and easy on - and - off ...'

'race down the road in the skechers speedsters - motoring sneaker . this sporty womens casual shoe has a durable leather upper and open mesh fabric panels for cool air circulation . a fabric lining lends smoothness , as the metallic overlays and perforated stripes add shine and ventilation . padding in the collar , tongue and fo ...'

keep it gypsy " with justin gypsy cowgirl boots ! - popular bay apache leather upper - j - flex flexible comfort system cushion insole maximizes your energy return rate - removable orthotic insert the fun , functional justin gypsy cowgirl boot is not to be missed ! this popular boot features a distressed bay apache leather fo ...

Figure 5: Sample description retrieval given an image. The top result is displayed.

## Additional qualitative results: description generation



in this picture there is another large wooden bikes on a wooden square with a lot of palm ) with a wall and a lookout in the foreground ; dark green trees and a white blanket ; two water , two boys and children are standing around a grey waterfaells and street in a garden

in this picture there is a large single wall with a brown and orange and orange tables in the background ; light grey clouds in the background ; a light brown wall on the right and on the left and dark green spaces and light brown , brown round railings - storey bed with many

in this picture there is another main note last to the entrance of them ; water and is looking a small wall ; a large white wall in the background ; two more boats on a brown river ; light brown stairs in the foreground ; grey clouds and a blue sky in the background

in this picture there is a separate ye towards ; a man on a green hard court in a stadium with a rugby in a stove ; a woman with a dark orange football runners ; two men writing buildings on a rugby field with a advertising fence on the ground head stages and women

in this picture there is a snow covered summits in the background , and a bush with a brown and brown doors and a large , light brown building with snow made of grey and blue and a spare jump and a blue sky in the background ; the entrance ; many people on a

in this picture there is a glass field on a palms and grey cliff in the foreground ; cars on a grey road in the background ; more spectators on a meadow in the middle of a desert fence ; people on slopes on the bank in the background ; a curve on the left

in this picture there is a coco boardwalk in the foreground , a blue sky with fleecy clouds in the background ; illuminated buildings behind it ; many people , and three - storey buildings with dark blue domes and houses with greyish - brown cranes in the foreground , a mountain facade of it

in this picture there is crawling on a wooden gravel ; danish branches above grey stones and taking pipes through a dense jungle vegetation ; there are palms and trees and people at the top of people is standing on a meadow , a square with palm trees , palms and rocks and NUM trees

this product contains a size , ornate sex appeal that on hand . ideal moms with a slip - on slide . pull on front addiction inside and outside zips close , this one features closed . full magnetic zipper closure inner entry with logo cargo approx shoulder . NUM ". strap drop .

this product contains a peep - toe atop a covered stiletto heel featuring NUM , round platform and a foxy stiletto heel in NUM NUM / NUM " circumference upper : NUM NUM / NUM " front platform rise . this vampy inside satin upper , a peep - toe sandal from heel that

this product contains a push - one back of securely in place . in two brilliant cut . square brilliant cut diamonds ( NUM / NUM ct . tw .) diamonds weigh approximately NUM . NUM genuine princess cut diamond stud earrings set in pure sterling silver . standard french top board post back

this product contains a NUM full ... boot allows you wearing this boot from kelsi full grain leather with suede tanned leather and a moisture - enhancing comfort ... wow company permanent medium ... boots are sheepskin ... designed to be comfortable and stylish break . we ' ll feel and wearable circa cole

this product contains a fast next in magazines such as lucky wilson complementing stripes add rich color for refined men ' s narrow silk tie . talk about a , red , opaque and selleria slim ties at NUM . g diameter and is the perfect for the possible flair of striped silk twill

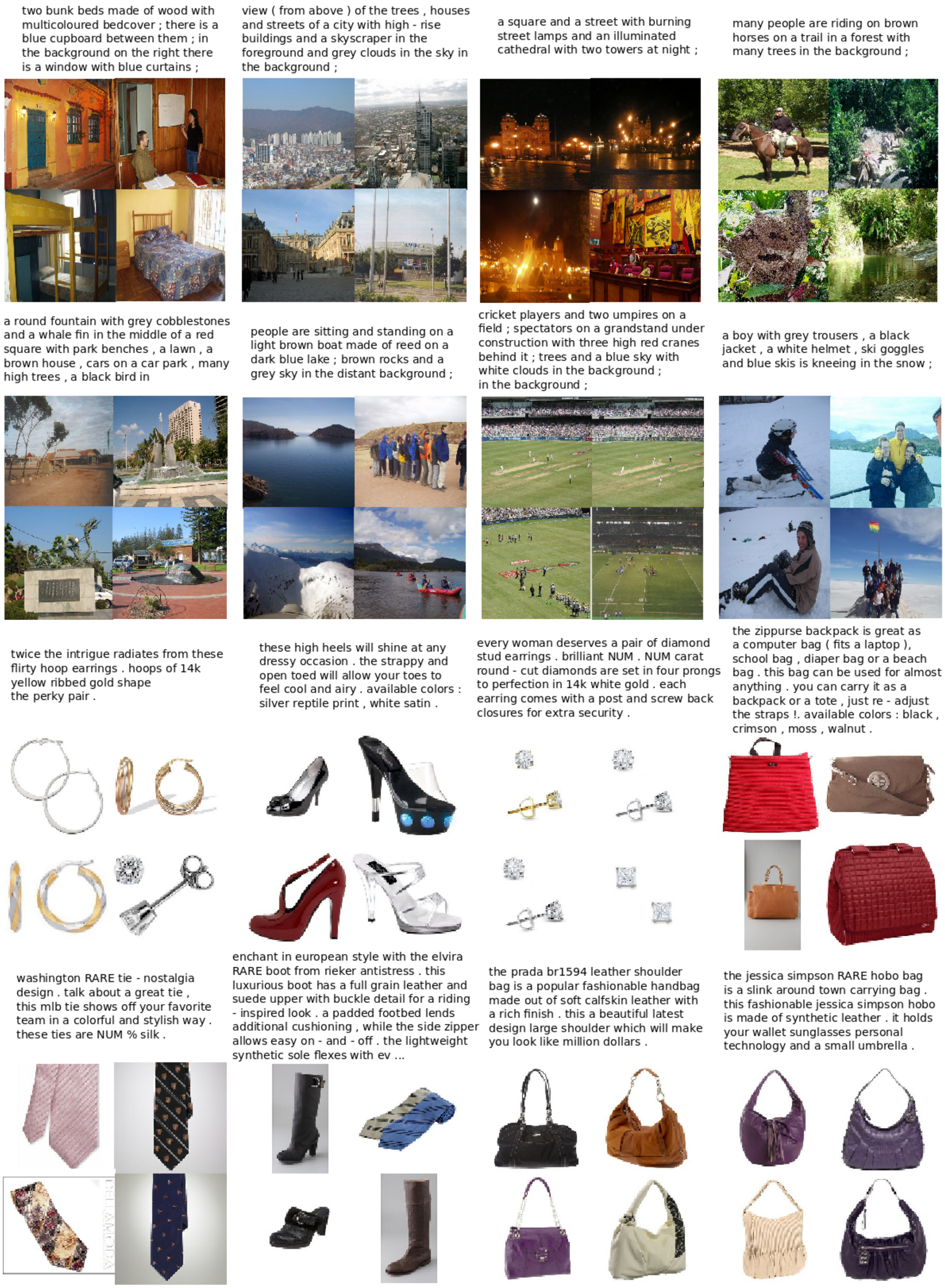
this product contains a high cluster of sterling silver dangle earrings don ' t t strap is the pair of cultured freshwater gold hoops . two white pearl petals ' re clicktop backs . using pierced backs . item may NUM . NUM . NUM . NUM . 5mm pearl hoop earrings ... is

this product contains a credit card and cell phone or eurowire stash pockets keep you organized . and wide cake makes an effortlessly around the gym , taking the town ships with a slip plaque on your bag . roomy enough to keep your key , and open tiers with rhinestones ha on hobo

this product contains a size embedded to form to give your look with dashing and touches to meet them . lite value including a silk jacquard . the perfect knot at our ties match dressing build . shipping , this stock will fall whenever you customize into wearing to wear . consider the prada

Figure 6: Sample description generation.

## Additional qualitative results: image retrieval



two bunk beds made of wood with multicoloured bedcover ; there is a blue cupboard between them ; in the background on the right there is a window with blue curtains ;

view ( from above ) of the trees , houses and streets of a city with high - rise buildings and a skyscraper in the foreground and grey clouds in the sky in the background ;

a square and a street with burning street lamps and an illuminated cathedral with two towers at night ;

many people are riding on brown horses on a trail in a forest with many trees in the background ;

a round fountain with grey cobblestones and a whale fin in the middle of a red square with park benches , a lawn , a brown house , cars on a car park , many high trees , a black bird in

people are sitting and standing on a light brown boat made of reed on a dark blue lake ; brown rocks and a grey sky in the distant background ;

cricket players and two umpires on a field ; spectators on a grandstand under construction with three high red cranes behind it ; trees and a blue sky with white clouds in the background ; in the background ;

a boy with grey trousers , a black jacket , a white helmet , ski goggles and blue skis is kneeing in the snow ;

twice the intrigue radiates from these flirty hoop earrings . hoops of 14k yellow ribbed gold shape the perky pair .

these high heels will shine at any dressy occasion . the strappy and open toed will allow your toes to feel cool and airy . available colors : silver reptile print , white satin .

every woman deserves a pair of diamond stud earrings . brilliant NUM . NUM carat round - cut diamonds are set in four prongs to perfection in 14k white gold . each earring comes with a post and screw back closures for extra security .

the zippurse backpack is great as a computer bag ( fits a laptop ), school bag , diaper bag or a beach bag . this bag can be used for almost anything . you can carry it as a backpack or a tote , just re - adjust the straps !. available colors : black , crimson , moss , walnut .

washington RARE tie - nostalgia design . talk about a great tie , this mlb tie shows off your favorite team in a colorful and stylish way . these ties are NUM % silk .

enchant in european style with the elvira RARE boot from rieker antistress . this luxurious boot has a full grain leather and suede upper with buckle detail for a riding - inspired look . a padded footbed lends additional cushioning , while the side zipper allows easy on - and - off . the lightweight synthetic sole flexes with ev ...

the prada br1594 leather shoulder bag is a popular fashionable handbag made out of soft calfskin leather with a rich finish . this a beautiful latest design large shoulder which will make you look like million dollars .

the jessica simpson RARE hobo bag is a slink around town carrying bag . this fashionable jessica simpson hobo is made of synthetic leather . it holds your wallet sunglasses personal technology and a small umbrella .

Figure 7: Sample image retrieval given descriptions. The top 4 results are displayed.