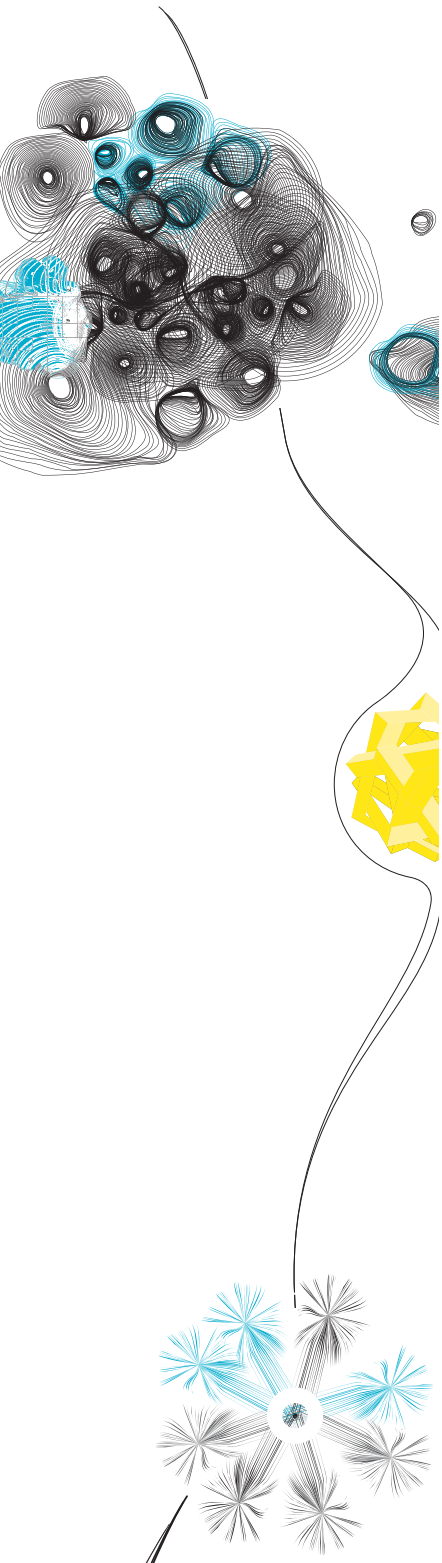
UNIVERSITY OF TWENTE.

Faculty of Electrical Engineering, Mathematics & Computer Science



Automatic Image Caption Generation for Digital Cultural Images Collections

Sanne van Waveren
Master Thesis
March 9th 2018

Graduation Committee:
Dr. G. Englebienne (University of Twente)
Dr. M. Poel (University of Twente)
Dr. S. Wang (OCLC)
R. Koopman (OCLC)

Human Media Interaction Group
Faculty of Electrical Engineering,
Mathematics and Computer Science
University of Twente
P.O. Box 217
7500 AE Enschede
The Netherlands

OCLC

Summary

Recent years have witnessed considerable growth of the volume of digital collections, which has lead to an increasing demand for automated techniques that support the management, navigation and search of these collections. As machine learning techniques are advancing, it becomes feasible to automatically generate image captions. However, one of the main challenges that needs to be addressed is to create captions that include higher-level information, such as the event or location shown in an image. Recently, the automatic image caption generation problem has been formulated as a translation problem. However, state-of-the-art on image captioning models captions are limited to low-level description of the image itself. In this work, we assume that images and text naturally co-occur and explore the feasibility of including title information in a pretrained state-of-the-art image captioning model using OCLCs CONTENT$_{dm}$ data. By using a combined objective function based on both title and image, we give a proof-of-concept for compressing image and title features with an autoencoder, so that can be used as input for a pretrained image caption generation model. Although the results are mixed, this thesis provides initial insights into the automatic generation of higher-level image captions.

Keywords: automatic image caption generation, historical digital image collections, long-short term memory, recurrent neural network, digital image collections, OCLC, CONTENT$_{dm}$

Contents

List of Figures

List of Tables

List of acronyms

ANN Artificial Neural Network

BPTT Backpropagation Through Time

CNN Convolutional Neural Network

DNN Deep Neural Network

DPL Denver Public Library

IIIF International Image Interoperability Framework

LSTM Long Short-Term Memory

MCRPL Manatee County River Public Library

MSCOCO Microsoft Common Objects in Context

NIC Neural Image Caption

NICAE Neural Image Caption + Autoencoder

OCLC Online Computer Library Center

PRPL Poudre River Public Library

RCNN Region-based Convolutional Neural Network

ReLU Rectified Linear Unit

RNN Recurrent Neural Network

SDSHSA South Dakota State Historical Society Archive

SGD Stochastic Gradient Descent

TF Tensorflow

TSV Tab-Separated Values

TPL Tacoma Public Library

W2V Word-2-Vec

Chapter 1

Introduction

Recent years have witnessed considerable growth of the volume of digital collections. This includes the digitisation of images and corresponding meta data. The growing popularity of online platforms, such as social media sites, online news sites, and in digital libraries provides large quantities of readily available digital images. As the volume of these digital collections continues to grow, there is an increasing demand for automated techniques that support both the management, navigation and search of these collections. Recent advances in machine learning, and deep learning in particular, have led to breakthroughs in object recognition and detection [1]. More specifically, the task of automatically describing images is increasingly receiving attention.

A quick glance at an image is enough for people to describe an image effortlessly [2]. It consists of roughly the following steps: 1) perception of the visual space, 2) conversion of the visual information into the language space, and 3) generation of a natural language description. In other words, it requires information from one modality (visual) to be translated into another (textual - either in spoken or written form). People are assumed to have an underlying representation for these two modalities [3]. However, a computer's semantic image understanding is less trivial. This discrepancy between people's high-level semantic image understanding and the low-level features extracted by computers, is one of the major challenges in automatic image caption generation [4, 5].

Although a handful of approaches have been proposed for automatic image caption generation, their reliance on hand-crafted designs makes them rather rigid in natural language generation [6]. Early work mostly focused on feature extraction and mapping them to a set of description words using a dictionary [5, 7]. As Feng and Lapata [8] point out, a major drawback of these approaches is their domain-specificity. Moreover, Héde et al. [7] and Feng et al. [9] argue that these approaches overlook the explicit relations between objects, which makes them ambiguous. For instance, "a female cat" explicitly indicates a relationship between "female" and "cat", while the words in isolation can also refer to two separate objects, a female and a cat.

More recently, Vinyals et al. [6, 10] suggest to treat automatic image caption generation as a translation problem. This approach combines a Convolutional Neural Network (CNN) for image processing and a Long Short-Term Memory (LSTM) recurrent neural network RNN for language modeling into a single joint model: the NIC model generates descriptions given an image only. Typically, machine translation models use an encoder-decoder architecture that encodes a sequence into a fixed length vector and decodes it into the desired output. In a similar fashion, an image can be "translated" into a caption and this approach has shown capable of generating reasonable captions that accurately describe an image's visual content. Specifically, the NIC model is good at describing objects and the relations between them. For instance, in a caption like "a child (object) is sitting (relation) on a chair (object)". One major advantage of the NIC model is that it requires only an image

to generate a caption, which makes it a knowledge-lean model.

However, as Bernardi et al. [11] point out, the work by Vinyals et al. (2015) also showed that differences in vocabulary and the quality of the captions substantially affect performance, and transfer learning between datasets is less effective. Hence, one of the main challenges that needs to be addressed is to generate more linguistically meaningful image captions. Up until now, image caption models are limited to creating captions for specific datasets and to captions that describe the image itself. Oftentimes, however, there is more information that is not available in the visual space. For instance, the location at which a picture is taken or the name of a person depicted in the image. Captions that include both information about the image itself and this "meta-information" can be especially interesting for digital collections of libraries and institutions. Those collections typically aim to provide rich background information to users. However, automatically generated captions may not easily scale up to the highly expressive language that can be found in people-generated captions [12].

In this work, we assume that images and textual information often naturally co-occur. Moreover, such textual information oftentimes describes the image on a higher level. For instance, imagine portraits of people for which it is highly likely that a description includes the name or status of the people portraited. Also, imagine images of events such as protests or parades for which it is expected that the description states the specific event or date. For a dataset of historical images, we propose to combine features from two modalities (i.e., images and image titles) and feed it as input to the NIC model. The novelty of this work is in the combination of image and textual information as input to a LSTM language model and in the type of data that we use; we use images of relatively varying quality and a large lexical diversity. Although using title information makes the model less knowledge-lean, it is interesting to explore the possibilities of using readily available meta-data.

1.1 Project Context

This project is carried out for the Online Computer Library Center (OCLC)[1] in Leiden, the Netherlands. OCLC is a non-profit cooperative organisation that offers technological services to support libraries in making their information accessible and useful for people all over the world. Another important aspect offered by OCLC is innovative research to help libraries to fullfil the ever-changing needs of their users. Driven by this motivation to support knowledge sharing, OCLC adheres to the motto: '*because what is known must be shared*'.

One of the services delivered by OCLC is their CONTENT$_{dm}$[2] system. This is a digital collection management system that enables users to build, preserve, and showcase their digital collections. It supports collections for over two thousand organizations world wide. Users can build their own digital collections with CONTENT$_{dm}$ and showcase these on their website. The CONTENT$_{dm}$ system's database contains roughly 40 million records. In this work, we focus on the image data, and the corresponding meta-data like the images' file name and format. The only required field when creating a new collection is the title, other meta-data fields are optional. As a result, the amount of the available meta-data varies across collections. A more detailed description of the data is provided in Section 4.2.

The main goal of this thesis is to explore the feasibility of creating captions using a pretrained state-of-the-art image caption generation model (i.e., NIC) using image title information. The motivation behind this exploration is that it can be useful to automatically create captions that include

[1] **www.oclc.org**

[2] www.oclc.org/en/contentdm.html

higher-level “meta-information” about the image. From this, the following research questions are derived:

- How can title information be used in a pretrained NIC model for OCLC’s CONTENT$_{dm}$ images.
- What effect does the use of title information have on the caption generation?

1.2 Report organization

The rest of this thesis is organised as follows. A background on the concepts used throughout this thesis is given in Chapter 2. Chapter 3 discusses prior work on relevant research directions, including but not limited to image annotation and object recognition. Chapter 4 introduces the methodology and framework used in this work. Chapter 5 discusses the experiments and the results. The results and suggestions for future work are discussed in Chapter 6. Finally, this work is concluded in Chapter 7.

Chapter 2

Background

This chapter provides a background on Artificial Neural Networks (ANNs), and it discusses the concepts of RNNs and CNNs in particular. The aim of this chapter is to create a better understanding of the models that are discussed throughout this work. First, Section 2.1 briefly introduces the concept of ANNs. Then, Section 2.2 discusses RNNs, which are a special variant of neural networks that introduce memory and it focuses on a particular variant of RNN: the LSTM. Finally, Section 2.3 describes CNNs, commonly used in computer vision for object detection and recognition. This chapter serves as background knowledge to better understand the related work discussed in Chapter 3 and the work presented in this thesis.

2.1 Artificial Neural Networks

ANNs are motivated by the idea that computations in the biological brain differ than computation done by conventional digital computers. They are inspired by networks of biological neurons in the following way: layers of simple computing nodes (similar to neurons) operate as nonlinear summing devices, which are interconnected through weighted connections (similar to synaptic links) [13, 14].

As early as 1943, the mathematical representation of neural networks has been an important focus of research and development [15]. Special variants have been designed to learn sequential, time-dependent patterns [16]. Neural networks have the potential for solving a range of function approximation problems; with the appropriate parameters, it has been shown that feed-forward neural networks can approximate any continuous function $f^* : \mathbb{R} \rightarrow \mathbb{R}^d$ to any desired degree of accuracy [17].

Figure 2.1 illustrates a simple perceptron with n input neurons and one output neuron. This multiple linear regression model is denoted as follows:

$$y = \sum_{i=0}^{N} w_i x_i. \tag{2.1}$$

In case of this simple perceptron in (2.1), the compute nodes x_0, x_2, ..., x_n in the **input layer** are linked with the compute node y_0 in the **output layer** through weighted connections w_0, w_1, ..., w_n (the **weights** of the model). The x_0 compute node is called a **bias unit** and incorporated in the summation by w_0. Typically, the output y_0 is transformed using a differentiable, nonlinear activation function like a hyperbolic tangent or logistic function [18, 19]. Then, we write $a_i^{(l)}$ to denote the activation of unit i

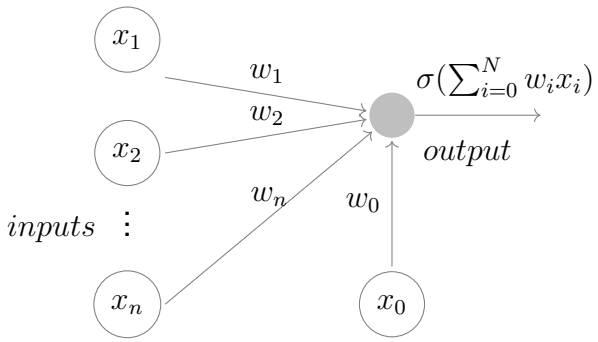


Figure 2.1: A simple perceptron with n input neurons and one output neuron. The output is transformed using a logistic activation function σ.

in layer l for $l \in 1, ..., N$, in case of our example in Equation 2.1, this gives:

$$a_1^{(2)} = \sigma(\sum_{i=0}^{N} w_i^{(1)} x_i^{(1)}), \tag{2.2}$$

where the total weighted sum of inputs can be denoted as z_i^l, which simplifies the notation of Equation 2.2 to:

$$a_1^{(2)} = \sigma(z_1^2), \tag{2.3}$$

More generally, the layer l's activations a_{l+1} are computed as follows:

$$a^{(l)} = f(z^l). \tag{2.4}$$

In theory, the number of nodes and connections is not restricted and networks can become highly complex. Besides an input and output layer, neural networks often comprise **hidden layers**. Traditional ANNs have one or two hidden layers, a model with more hidden layers is considered a deep neural network. The deeper a model, the more data is required for training. The hidden layer processes the actual output. In other words, this is where learning takes place; it gets its inputs from the input or previous hidden layer and learns to match this to a certain output. How this learning takes place is discussed next.

2.1.1 Learning Approaches

Four major machine learning approaches can be distinguished: supervised, semi-supervised, unsupervised, and reinforcement learning. The data, the availability of labeled data in particular, determines the appropriate type of learning. Below, the approaches are discussed.

Firstly, supervised learning requires all data to be labeled, since it refers to the task of inferring outputs from labeled training data. For instance, if a model learns to classify whether a picture depicts a cat or a dog by looking at pictures and their corresponding labels. Especially for more complex problems, this requires a large amount of labeled data, and consequently, a fair amount of work to create those labels if they do not already exist.

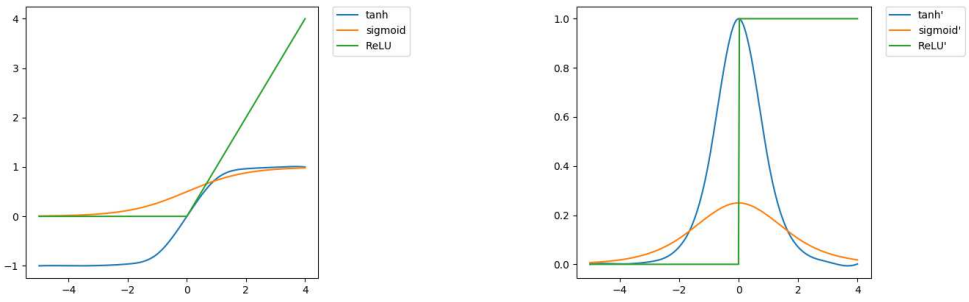
Secondly, semi-supervised learning combines supervised and unsupervised learning: data is divided into labeled and unlabeled data sets. Within semi-supervised learning there is a distinction between inductive and transductive learning [20]. The goal of inductive semi-supervised learning is to learn to predict future data better than when learning from labeled data solely. Transductive

learning is used when there is an interest to predict unlabeled training data, but there is no intention to generalize to future data.

Thirdly, unsupervised learing refers to training with unlabeled data only; the model tries to infer an unknown or hidden structure from unlabeled data. The main goal of unsupervised learning is to find hidden patterns in the data, because the model's output cannot directly be compared with the correct output due to the lack of labeled data.

Lastly, reinforcement learning is particularly useful in situations where the decision or action made by the model in a state that is dependent on the input data. Consider a model can classify a picture of a cat as '`cat`' at t_x. At a later moment in time t_{x+n}, this picture will still be a picture of a cat. Therefore, the only correct label to be learned is '`cat`' in this case. However, this is different when given a model that needs to learn to play a game of chess. A move that is considered to be good at time t_x does not have to be a good move at time $t_x + n$. The set of good moves is dependent on opponent's move that happened between time t_x and $t_x + n$.

Basically, the aim of training a ANN is to reduce a loss function to learn the optimal weight matrix **w** and bias vector **b**. The loss functions computes the difference between the model's output and the desired output. For instance, when classifying a picture of a cat, the model's output can be one out of N classes, whereas the desired output is the class *cat*. During training, the model tries to learn the correct output by changing the values of its weights and biases.



(a) *(b)*

Figure 2.2: The plots for the sigmoid, hyperbolic tangent, and ReLU activations functions (a) and their derivatives (b).

2.1.2 Backpropagation Algorithm

The Backpropagation algorithm is a gradient descent technique to minimize a network's error E, which is commonly used in supervised learning problems. After each forward pass through the network, the loss function's partial derivatives with respect to the network's weights are computed using the chain rule [21]. The negative of the gradient vector is the direction that minimizes the cost function.

Importantly, in deeper models this may cause the model to get stuck in a local minimum instead of a global minimum. As compared to using all the train samples, updating the model's weights after an iteration using a mini-batch of data like in Stochastic Gradient Descent (SGD) has been shown to lead to faster convergence and better solutions.

The initialization of the weights is important, because it contributes to how fast the model can learn. Consider a network with one hidden layer using a sigmoid activation function. The sigmoid function is a real function $s_c : \mathbb{R} \to (0, 1)$ defined as follows (See Figure 2.2a for the graphical representation):

$$y = \frac{1}{1 + e^{-x}} \tag{2.5}$$

The derivative of the sigmoid function is defined as follows (See Figure 2.2b):

$$\sigma(x) = \frac{e^x}{(1 + e^x)^2} = \sigma(x)(1 - \sigma(x)). \tag{2.6}$$

Now, for larger or negative values for x the sigmoid's derivative takes values close to zero. Naturally, it becomes apparent that bigger gradients lead to faster learning, whereas smaller gradients lead to slower learning. Hence, initialization of the weights influences the model's learning pace. Moreover, symmetry breaking is important, which means that the weights have to be initialized randomly rather than to the same value like zeros [22]. If all weights are initialized using the same value, all hidden layer units end up learning the same function of the input, since the activation is the weighted sum of inputs as shown in Equation 2.2.

The idea behind the Backpropagation algorithm is to first compute activations and output in a forward pass through the network and then propagate the errors backwards. The activations and output values resulting from the forward pass are used to compute the partial derivative of the loss function with respect to the weights, defined as:

$$\frac{\partial E}{\partial w_{ij}^k} = \frac{\partial E}{\partial a_j^k} \frac{\partial a_j^k}{\partial y_j^k} \frac{\partial y_j^k}{\partial w_{ij}^k}, \tag{2.7}$$

where y_j^k is the pre-activation value of node j in the k-th layer, which is the weighted sum of the inputs (see Equation 2.1), so that:

$$\frac{\partial y_j^k}{\partial w_{ij}^k} = a_i^{(k-1)}, \tag{2.8}$$

and the first two factors are also referred to as error term δ:

$$\delta_j^k \equiv \frac{\partial E}{\partial a_j^k} \frac{\partial a_j^k}{\partial y_j^k}. \tag{2.9}$$

The partial derivate of the error with respect to the weights is used to update the weights as follows:

$$w_{ij}^k = w_{ij}^k - \alpha \frac{\partial E}{\partial w_{ij}^k}, \tag{2.10}$$

where the first term is a learning constant that defines the step length of each iteration in the negative gradient direction, and the second factor is partial derivative of the loss with respect to that weight, as defined in Equation 2.7.

Thus, roughly, the steps in the Backpropagation algorithm are as follows:

- Feed-forward step through the network;
- Backpropagation of the error through the network;
- Update the weights.

First, perform a forward pass through the network computing all activations and output values for all the nodes in the network. Then, compute the error terms from the output values. Last, update the weights according to the partial derivative of the error with respect to the weights. This is repeated until the error value is sufficiently small.

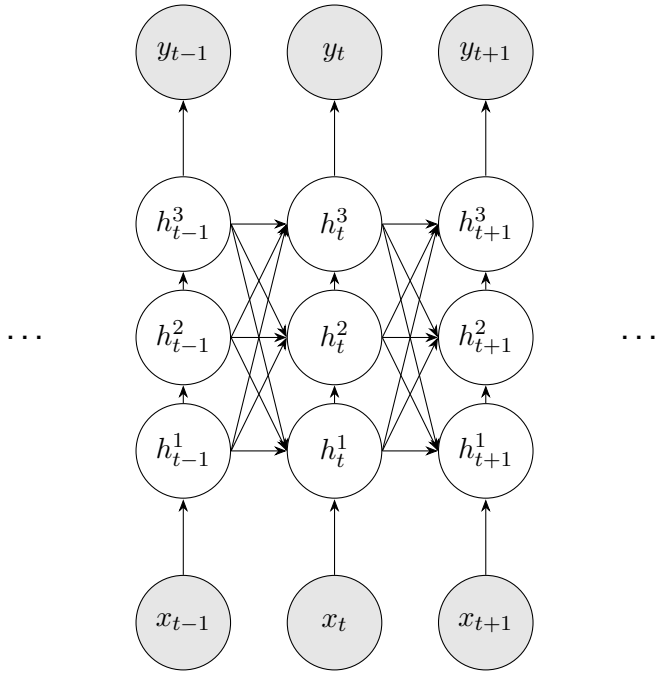


Figure 2.3: A three-layer RNN.

2.2 Recurrent Neural Networks

As you are reading this, your understanding of the current word depends for a large part on previous words. Similarly, context is crucial to computational language modeling, speech recognition systems and video-to-text translation [23, 24]. Though, such context is not incorporated in standard, vanilla neural networks, as their context is limited to $N - 1$ words [25]. RNNs address this problem by allowing cyclical, or recurrent, connections [26]. Hence, RNNs are a broad family of neural networks that have the capacity to carry over information from previous time steps by including the hidden layer h^l_{t-1} in the computation of h^l_t, where $l \in L$ and L is the network's depth, and $t \in T$ and T is the time (See Figure 2.3). The weights of the network can vary between layers, but are shared across time.

Recently, RNNs have emerged as effective models in situations that involve sequential data [27]. Applications include speech recognition [28, 29], hand writing recognition and generation [30, 31], machine translation [32, 33, 34], language modeling [35], and image captioning [6, 36].

More formally, as defined in Graves et al. [28], given an input sequence $x = (x_1, ..., x_T)$, a standard RNN computes its hidden layer h_t and output y_t for $1 \leq t \leq T$ as follows:

$$h_t = \mathcal{H}(W_{xh}x_t + W_{hh}h_{t-1} + b_h) \tag{2.11}$$

$$y_t = W_{ht}h_t + b_y \tag{2.12}$$

where W terms denote weight matrices, b terms denote bias vectors, and $\mathcal{H}$ denotes the hidden layer function like a sigmoid or hyperbolic tangent.

2.2.1 Vanishing and Exploding Gradients

Despite its recurrent nature, the RNN's architecture still poses limitations in terms of what it can capture over time (e.g., memory). Conceptually, a standard RNN can take into account all preceding words [37]; information can be carried over indefinitely. However, in practice, it has been shown that

gradient-based Backpropagation Through Time (BPTT) (for a more elaborate discussion of BPTT see Werbos [38]) causes gradients to vanish or to explode [26, 27, 39, 40]. As can be seen in Figure 2.2, the derivatives for the sigmoid, hyperbolic tangent activation functions are almost always smaller than one, and hence, we are multiplying a lot of small numbers together. This can also happen with a large amount of big gradients multiplied together, resulting in large terms. These two problems, also known as the *vanishing* and *exploding gradient problem*, causes the network to be unable to learn long-term temporal relations. Exploding gradients can relatively easily be solved by truncating or squashing, but vanishing gradients is more difficult to solve. However, with the Rectified Linear Unit (ReLU) activation function defined as

$$y_t(x) = \max(0, x) \tag{2.13}$$

the derivative is zero when the pre-activation term is less than zero, and one otherwise, which prevents the gradients from vanishing. Moreover, we can initialize our weights by an identity function rather than draw them from a standard normal distribution. Additionally, the vanishing gradient problem is addressed by a special kind of recurrent neural network, that uses gated cells. One variant of a gated cell is called Long Short-Term Memory (LSTM), which is discussed in more detail next.

2.2.2 Long Short-Term Memory

LSTM networks, as introduced by Hochreiter and Schmidhuber [41], do not have a fundamentally different architecture from RNNs, however, they can learn long-term dependencies without having the gradients to vanish or explode like with standard RNNs. LSTMs are a type of recurrent units that use logical gates to control the flow of information that goes through. It uses addition to determine the subsequent state instead of multiplication, and this seemingly simple change leaves the error more constant.

Instead of the summation units in the hidden layer of a standard RNN, LSTM networks make use of memory blocks [26]. These memory blocks contain at least one memory cell and three multiplicative units: the input, forget, and output gates (See Figure 2.4). The gates make it possible for the LSTM memory cells to decide what to keep in memory and what information to erase over long periods of time. The equations for the gates and cell state are as follows:

$$f_t = \sigma(W_f\dot{[}h_{t-1}, x_t] + b_f) \tag{2.14}$$

$$i_t = \sigma(W_i\dot{[}h_{t-1}, x_t] + b_i) \tag{2.15}$$

$$o_t = \sigma(W_o\dot{[}h_{t-1}, x_t] + b_0) \tag{2.16}$$

$$C_t = f_t * c_{t-1} + i_t * h(W_C\dot{[}h_{t-1}, x_t] + b_C) \tag{2.17}$$

Roughly, the forget gate decides what information is irrelevant and can be disregarded. Consider the case that we are modeling language, if the input contains a new subject we can choose to forget the previous subject, because it is likely that the next word will be conjugated according to the new subject. For instance, when the subject first was “my dog” and the new subject is “my parents and I”, or when the gender of the subject changes, such as with the old subject being “the girl” and the next subject being “the old man”. The next step is to selectively update the cell state values, which is done by the input gate and an update layer (e.g., hyperbolic tangent layer) that creates the new

candidate values for the cell state. In the last step, the output gate decides what is going to be the actual output. For instance, by applying a sigmoid layer on the cell state values to decide which values are being kept, pushing the cell state values between -1 and 1, and then multiplying this by the sigmoidal output gate.

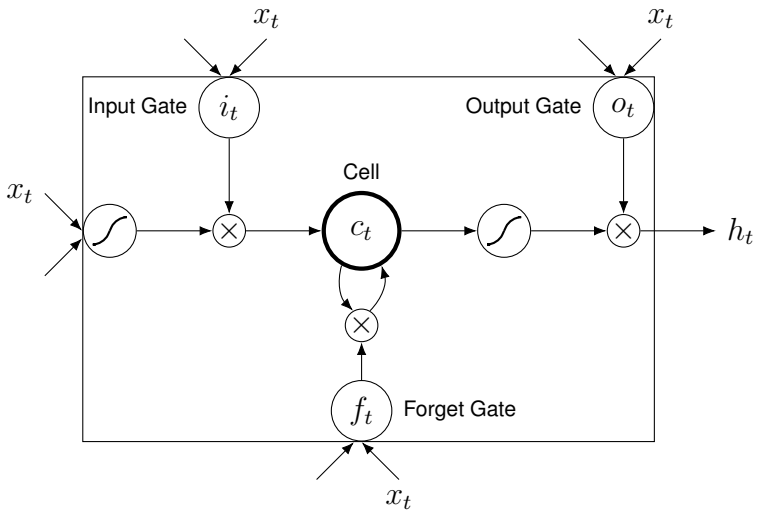


Figure 2.4: Long Short-term Memory Cell.

2.3 Convolutional Neural Networks

Given an image dataset CNNs are commonly used as they are inspired by the idea of translational invariance. This means that these models can detect objects independent of where it is located in the image, which is important given the high variance within images. Essentially, every image can be represented as a matrix of pixel values ranging from 0-255. A grayscale image consists of one two-dimensional matrix, whereas regular digital images have three two-dimensional matrices stacked upon each other, corresponding to the three colors: red, green, and blue. CNNs incorporate the notion of local features that depend on a sub region of an image, based on the assumption that surrounding pixels have a stronger correlation than more distant pixels [18].

The notion of local features is incorporated into CNNs through the following mechanisms: local receptive fields, weight sharing, and subsampling [18]. At the basis, the building blocks of a CNN are the following four: convolution, non-linearity, pooling or subsampling, and classification.

The convolution of a five-by-five image and three-by-three kernel or feature detector, means that the three-by-three matrix slides over the five-by-five image matrix by a certain number of pixels, this number is called stride. For every position, the sum of the outputs of an element-wise multiplication between the two matrices forms a single value of an output matrix, also called a feature map. Since weights are shared between units in the feature map, the network outputs are invariant to translations and distortions of the image [18]. Though, it is important to note that parameter sharing does not make sense if it is expected that for one part of the image completely different features have to be learned than for another part. For instance, for face detection with faces centered in the image. Next, introducing non-linear functions like ReLU after the convolutional layer enables us to account for non-linearity, since most real-world data is expected to be non-linear, but the previously described convolution is a linear operation. Commonly, a convolutional layer is followed by a sub-sampling or

pooling layer that performs down-sampling along the spatial dimensions; by reducing the number of parameters this layer controls for overfitting. Sub-sampling or pooling means that another window slides over the convolved feature, and aggregates these values into one value (e.g., the maximum in max pooling and the mean in mean pooling). Finally, the last layer of a CNN is a fully-connected layer for classification. For instance, using a softmax classifier in which the output probabilities sum up to one, and the highest probability is the most likely class (e.g., given a picture of a dog the model is expected to output the highest probability for the class that represents the category dog).

CNNs have proven very effective in many computer vision tasks, such as face recognition, image classification, and age and gender detection [42, 43, 44]. For a more elaborate discussion of CNNs, see Thoma [45], and for the use of CNNs for image classification in particular, see Krizhevsky et al. [43] and Simonyan and Zisserman [46].

Chapter 3

Related Work

A quick glance at an image is sufficient for most people to describe an image effortlessly [2]. It has been reasoned that this is because people have an underlying representation for the two modalities [3]. For a computer, on the other hand, such semantic understanding of a scene is less trivial. Generally, the image description process can be divided into the following parts: 1) perception of the visual space, 2) convert this visual information using world knowledge into the language space, 3) generate textual description [47]. The major challenge for tasks like image annotation and image captioning is the discrepancy between people's high-level interpretations of image semantics and the low-level features derived by a computer, also known as the semantic gap [4, 5].

Hence, automatic image caption generation is a non-trivial task for a computer and only a decade ago it was still deemed impossible. It requires the computer to both recognize and localize objects, and to describe the relation between them. However, recent advancements in computer vision have proven this complex task possible and this has led to an increase in work done on automatic image caption generation. To provide a better understanding of these developments, this chapter discusses prior work on image annotation, and image caption generation.

3.1 Image Annotation

The rapid growth of online data has inherently increased the demand for algorithms that can search and browse large-scale image collections. However, many search engines rely on keywords to retrieve images and this poses a problem when image annotations are absent. As a result, there has been an increasing demand for automatic image annotation. A formal problem formulation for image annotation is given by Feng and Lapata (p. 23) [8]: *"Given an image I with visual features $VI = v_1, v_2, ..., v_N$ and a concept set $W = w_1, w_2, ..., w_M$, where M is the number of concepts, the image annotation task is to find the subset $WI(WI \subseteq W)$, which can appropriately describe the image I"*. Essentially, given an image, the goal is to assign relevant labels based on visual features [48]. One major problem in image annotation is the availability of training data, as correctly labeled data requires human labor, and the diversity of real world images requires the amount of training data to be large [8]. Early work typically used labels created by domain experts or weak labels such as user-generated tags, which are commonly found on social network sites, and search engine query terms [49, 50].

Previous work in image annotation can be divided into two streams: the study of isolated object detection and the study of the semantic relation between them [5, 8]. For instance, the former focuses on objects such as *person*, *street*, *sidewalk*, and the latter focuses on the hierarchical organization of

those concept. Fan et al. [4] for instance, use a concept ontology to organize the concepts hierarchically, in which *sidewalk* is subcategory of *street* [4]. This work lies at the basis of automatic image caption generation. Similarly, closely related to image annotation, object recognition can be seen as an important subtask in the context of image captioning and is discussed next.

3.2 Object Recognition

Object recognition is the task of detecting a given object present in an image or video [51]. It has received great attention from the field of computer vision; this is reflected in large scale, yearly competitions [52]. High variability within classes, like a wide variety of lightings, corruptions, occlusions, and so on, make image classification a challenging task [53]. Visual perception requires an internal representation of the world or aspects ("features") of that world, that is invariant to irrelevant variations of the input [54]. Many attempts have been made to manually design low-level features for classification, and more recently, there has been an interest in methods to learn features directly from data. Especially, Deep Neural Networks (DNNs) in combination with convolutional architectures have proven successful in image classification [54]. As a result, object detection is computationally expensive, though, recent advances with fast Region-based Convolutional Neural Networks (RCNNs) have achieved near real-time rates using very deep networks [55].

CNNs have become popular in large scale image recognition tasks [43]. Several deep networks were proposed, such as Network In Network by [56], VGGNet by [46], and GoogLeNet by [57]. More recently, residual connections were introduced, the use of these connections improve training speed of those very deep networks greatly [58]. The GoogLeNet, or Inception model, was revised several times, with the most recent architecture being Inception-v4 [59]. While object recognition and image classification techniques have been well-studied, the task of automatic image caption generation is significantly more complex [10]. The next section discusses prior work on image caption generation is discussed.

3.3 Automatic Image Caption Generation

The ability to automatically describe the content of images has several applications, including but not limited to helping visually impaired people better understand visual content [8, 60], assisting people in generating a image captions [61], and to support more accurate and targeted queries for end users of image search engines [8, 9].

Automatically describing an image in natural language is a complex task, which combines two fields that have typically been treated in isolation: computer vision and natural language processing. A handful of approaches automatically generate image descriptions using a two-stage architecture. The image is represented by image features (e.g., color, detected edges) which are converted into an abstract representation, which is then rendered into a natural language description with a text generation engine.

For instance, Kojima et al. [62] propose a method to create human action descriptions in office scenes using body motion features (e.g., head position and direction) and concept hierarchies of actions. Héde et al. [7] use pictures of objects with uniform backgrounds. The system relies on a manually created dictionary of objects indexed by an image signature (e.g., color and texture) and two keywords (object's name and category). First, image processing segments images into objects, their signature is created and compared to signatures in the dictionary. Then, a text description is

generated using the keywords associated to the images of the database. Similarly, Yao et al. [5] use an image parser and visual knowledge base to generate text descriptions for images and video. A parse graph includes contents of the scene (e.g., objects and labels) and additionally, the graph includes horizontal directions for specifying relationships and boundary sharing. These hierarchically decomposed images are converted into a semantic representation using a lexical semantic network, which are then converted into human readable text using a text generation engine. Again, a drawback of this approach is the reliance on a large-scale manually created ground truth image database. Farhadi et al. [63] evaluate the similarity for an image-sentence pair by mapping them both to a representation the meanings space and compare results. This representation is a triplet of scene elements that is converted to text using templates. Similarly, Li et al. [64] compose sentences from scratch using detections, from which meaning representations are constructed and then put together phrases that contain the detected objects and scenes.

As Feng and Lapata [8] point out, the major drawback of these approaches is that due to their reliance on hand-crafted knowledge, they are domain-specific and limited to specific concepts or scenes. The mapping dictionaries are usually created manually. Moreover, the work by Héde et al. [7] illustrates the importance of explicit relations between objects. For instance, “a female cat” explicitly indicates a relationship between “female” and “cat”, while the words in isolation can also refer to two objects, a female and a cat.

Recent state-of-the-art models use captions directly in training. For instance, Fang et al. [65] propose a system that trains on images and their captions, and learns to extract word forms (verbs, nouns, adjectives) from regions in the image. These words then guide a language model to generate human readable captions, which are then re-ranked using a Deep Multimodel Similarity Model (DMSM).

Interestingly, Vinyals et al. [6, 10] formulate the automatic image caption generation task as a translation problem. The NIC model combines a CNN for image annotation and a RNN for sequence modeling into a single network that can generate descriptions given an image. Work in statistical machine translation has shown that directly maximizing the probability of the correct translation given an input sentence in an end-to-end fashion achieves state-of-the-art results [66]. Since these models make use of an encoder-decoder architecture that encodes a sequence into a fixed length vector and decodes it into the desired output, Vinyals et al. [6, 10] reason that is only natural to formulate automatic image caption generation in a similar fashion, where the image is “translated” into the desired caption. The initial state of an LSTM is computed by feeding it an image (I_0), whereafter the words (S_0, ..., S_{n-1}) of the true description are fed sequentially to the LSTM. Inputs are projected into the same vector space using a vision CNN for the image and word embedding W_e for the words. Their results showed that the model is capable of generating reasonable captions that describe the objects present in an image and the relation between them.

Chapter 4

Methodology

The problem formulation determines the approach we take to answer our research questions. For this reason, Section 4.1 discusses how we treat the task of automatic image caption generation. Moreover, the data and is described in more detail in Section 4.2. In this work, we make use of Tensorflow (TF)[1], Google's machine learning library, which is introduced in Section 4.3.. Then, the data preprocessing steps are discussed in Section 4.4. Additionally, the NIC model that is used for generating image captions are introduced in Section 4.5. Finally, the approach to incorporating title information is described in Section 4.6.

4.1 Problem Formulation

In line with Feng and Lapata [9], we formulate image caption generation as follows. Given an image I, and related knowledge base K, generate a natural language description C that captures the main content of the image under K. Specifically, for the CONTENT$_{dm}$ data, this means that data consists of (image,title,caption) tuples (See Figure 4.1). In this work, we approach the image caption generation as a translation problem similar to Vinyals et al. [6]. This means that we use an LSTM language model as decoder to output a sequence of caption words. However, our approach differs from Vinyals et al. [6] in that we combine the image features obtained through a CNN decoder with title information from the knowledge base K.

We focus on exploring whether this model architecture is capable of creating captions that contain both information about the image itself (e.g., the objects in it and the relations between them) and "meta-information" (e.g., names, events, and other higher-level information). The latter type of information is generally not extracted from the visual information, but it may be conveyed in textual meta-data that often naturally co-occurs with images. This is especially interesting for digital collections of libraries and institutions, which typically aim to provide rich background information to users. Though, automatically generated captions may not easily scale up to the highly expressive language used in people-generated captions [12].

4.2 Data

The CONTENT$_{dm}$ system contains roughly forty million records that consist of audio, visual, and textual data. In this work, we only include the set of entries that consist of image and textual data.

[1] tensorflow.org

Caption: Men chop wood to fuel the fires under large, kettle stoves used to cook food for the cast and crew of Buffalo Bill's Wild West Show.
Title: Wild West Show kitchen crew

Caption: New machinery at Titus Manufacturing Company, Mr. Leon Titus. A workman is bench testing rebuilt carburetors.
Title: Richards Studio D43896-12

Figure 4.1: Two example data entries. Each data entry during training consist of an image, caption and a meta-data (e.g., titles).

Several different sources like libraries and institutions have contributed in the creation of the collections.

Five collections of in total over 200K historical images have been selected based on their ostensible similarity. The reason for selecting collections that appear to be similar to each other, is to increase the chance that our model can learn at least some pattern. However, it is not straightforward to objectively determine the semantic similarity between large collections of images, hence, we determined similarity of the collections based on a visual inspection of an arbitrarily selected subset of the images. The following collections are included: Denver Public Library (DPL), Manatee County River Public Library (MCRPL), Poudre River Public Library (PRPL), South Dakota State Historical Society Archive (SDSHSA), and Tacoma Public Library (TPL).

Table 4.1 reports the number of images in every collection, how many images have a textual description, the maximum and average length of the description, and the number of unique words in all descriptions (e.g., vocabulary size). All collections contain historical images that are most, if

Collection	Images	Descriptions	Maximum	Average	Vocabulary
DPL	90,131	88,742	517	30	44,325
MCRPL	33,312	32,926	592	20	26,791
PRPL	33,212	31,833	528	18	23,311
SDSHSA	64,566	64,189	227	21	20,202
TPL	36,446	36,111	368	67	70,826
Total	257,667	253,801	446	31	185,455

Table 4.1: The five selected collections of historical images and some descriptives.

Figure 4.2: MSCOCO image retrieved on February 16th 2018 from `http://cocodataset.org/#explore?id=177529` and one of its reference captions.
Caption: a little kitten sitting on a stool next to a big brown dog

Figure 4.3: MSCOCO image retrieved on February 16th 2018 from `http://cocodataset.org/#explore?id=521838` and one of its reference captions.
Caption: a red train engine traveling down tracks near a forest.

not all, related to American history. The title is the only obligatory meta-data field, which means that for all images there is a title. Optional fields can include among others a description, date, technical details, and additional notes. The five collections almost all contain captions, hence, we can extract a train dataset of sufficient size without additional work to manually create captions. This allows for supervised learning, meaning that the NIC model makes use of the correct captions during training.

The $\text{CONTENT}_{\text{dm}}$ image captions are higher-level image descriptions as compared to the more general Microsoft Common Objects in Context (MSCOCO) image captions. We illustrate this with two examples of images and their corresponding captions from the MSCOCO dataset in Figure 4.2 and Figure 4.3.

4.3 Tensorflow

The Tensorflow (TF) open-source software library for machine learning is used as implementation framework. The choice for this library is based on its online support community and its flexibility. TF has a highly modular architecture, which means that parts can be created and used independently. Moreover, TF has a feature called Tensorboard, to visualize training and evaluation of models. This helps providing insights into the training process. Additionally, TF is portable and has advanced support for among others, threads and queues, which are important features when working with large datasets commonly used for image caption generation. Finally, we use an implementation of the image caption model by Google built in TF, and hence, it is a convenient choice to use the same framework for any modifications.

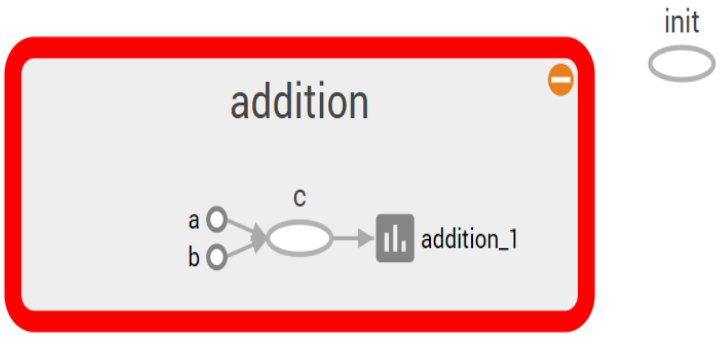


Figure 4.4: A simple Tensorflow graph as shown in Tensorboard, which contains two inputs variables (`a` and `b`) and an addition operation (`c`), and an `init` operation for the initialization of the variables.

TF requires computations to be implemented as data flow graphs, which need to be created on beforehand in TF. So called placeholders (`tf.Placeholder()`) can be specified for input, which are fed data at runtime. Nodes in the graph represent mathematical operations (e.g., `tf.matmul`), while the edges represent multidimensional data arrays called tensors. Figure 4.4 shows an example of a simple TF graph visualized through Tensorboard, which contains two inputs `a` and `b` and an addition operation `c`. The `init` operation represents an initialization operation, which initializes all variables in the graph.

Sessions are an important concept in TF, because they are necessary to actually run any data through the graph. After specifying the graph structure, it is necessary to initialize all variables and then run them. Otherwise, if you try to access a variable it will only return the tensor, and not its actual value. So, in case of our example in Figure 4.4, if we want to evaluate any variable (`a`, `b`, `c`), we need to initialize a `tf.Session()` and call `sess.run(v)`, where `v` is the specific variable we want to evaluate.

The graph and evaluated variables (e.g., loss) can be visualized in Tensorboard through the following steps. For any desired variable we create a `tf.Summary`, which can store scalars, histograms (e.g., weights and biases matrices), images, audio, and text. Once we created summaries for all variables, we merge them (`tf.summary.merge_all()`). It is important to merge them, because technically the summary nodes are not run, neither does their output depend on an operation in our graph: the summary operation is not run. Only operation that are run in TF produce output, and hence, we need to run all summaries. To avoid a tedious task, we combine all summaries and run them once. Then, we create a writer (`tf.summary.FileWriter`), which we provide with a directory to log the summaries to and the current graph (`sess.graph`), and we use it to write the summaries to the log directory. Importantly, to view the actual summaries in Tensorboard, you may want to add `writer.flush()`. Finally, we can run Tensorboard providing it with the log directory that now contains the freshly generated summaries.

4.3.1 Input Pipeline

TF has a specific input pipeline for fetching data files. There are three ways to provide data to a TF graph:

- Preloaded data: a variable or constant in the graph holds all data;
- Feeding: Python code provides data for each step;
- Reading from files: an input pipeline reads data from files at runtime.

The first two methods are typically used when data fits into memory. Importantly, the difference between preloaded data and feeding is that with the former all data is preloaded into variables or constants (as in the example in Figure 4.4) and with the latter Python code (e.g., a generator) creates data at runtime and feeds it to the TF graph. Usually, this in done by creating a `tf.Placeholder` and feeding data through a `feed_dict`.

Reading data from files is preferable when working with large amounts of data that do not easily fit in memory. For this, the data needs to be in a specific file format called `TFRecord` files, which contain `SequenceExample` protocol buffers. An advantage of this method is reusability, because reusing a model requires to convert the data into `tf.SequenceExample` format, without changing the actual model code. An `Example` is a data structure that consists of a `context` for non-sequential features and `feature_lists` for sequential data. The Feature values can be a `BytesList`, `FloatList`, or `Int64List`. The `SequenceExample` also contains a context which contains non-sequential features which apply to the entire example. In case of our image caption data an example of a `SequenceExample` is:

```
context: {
   feature: {
      key: ``image_data''
         value: { bytes_list: {
                value: [encoded_image]
            }}
      }
   feature: {
      key: ``image_id''
         value: { float_list: {
                value: [1.0]
            }}
      }
   }
feature_lists: {
   feature_list {
      key: ``caption''
         value: { bytes_list {
            value: [``A dog on the beach'']
            }}
      }
   }
```

Given a list of filenames, the files are put into an input queue that outputs strings (e.g., filenames) to a queue. Then, a `tf.RecordReader` reads the files given this queue, and the results from the reading operation are stored into a values queue like `tf.FIFOQueue` (First-In-First-Out). The values in this queue are then used to retrieve a serialized `SequenceExample` and a `tf.parse_single_sequence_example` decodes the `SequenceExample` protocol buffers into tensors, which can be used to replace the `tf.Placeholders` in the graph.

4.4 Data Preprocessing

This section describes the steps that are taken to preprocess the data. As mentioned in Section 4.2, the data consists of images, descriptions, and titles (and additional meta-data, such as date, location, and so on). The data comes in Tab-Separated Values (TSV) files and one field contains the CONTENT_{dm} item URLs that points towards the webpage for that corresponding data entry. CONTENT_{dm} has adopted the International Image Interoperability Framework (IIIF) APIs for serving images, and by parsing information out of the CONTENT_{dm} item URLs, we created the IIIF API URL to retrieve the full images using the `wget` command on the OCLC's Ubuntu 4.8 Linux server.

Having the full images and textual data, we then created JSON files, because that is the file format that is expected by the script that later on generates the `tf.RecordFiles`. A Python script reads the data and creates annotations for all images: it stores all data entries that have a caption in a train file and all entries without caption in a test file. It is important to note that the textual data differs in layout per collection, which means that the captions for each image can be stored under different headings in the TSV file. Consequently, we had to specify some code to find the right column to properly extract the captions, titles, and image identifiers. The image identifiers are preceded by `'908070'` and an integer (0 = DPL, 1 = MCRPL, 2 = PRPL, 3 = SDSHSA, 4 = TPL) to indicate the collection the image belongs to. For instance, image with identifier 0 from the DPL collection has the image identifier `'9080700'`. The annotations are stored in a dictionary with the following format:

```
dictionary {
   key: 'images' {
      key: 'image_file' {
         value: 'full path to JPEG image file'
         }
      key: 'id' {
         value: integer
         } }
   key: 'annotations' {
      key: 'image_id' {
         value: integer
         }
      key: 'caption' {
         value: 'caption (unicode string) corresponding to the image'
         }       key: 'title' {
         value: 'title (unicode string) corresponding to the image'
         } }
   }
```

A second Python script builds the actual `tf.RecordFiles` that are used later in the actual models. It obtains all the data that is needed and stores them in `SequenceExamples`. The script uses multi-threading, and stores 90% of the train data (from the train JSON file) as train set, 10% of the train data as validation data, and 10% of the data (from the test JSON file, if needed, supplemented with train data) is stored as test data.

4.4.1 Text Preprocessing

Simple tokenization is performed on the textual data. Captions are converted to lower case, and we keep the words that occur at least 10 times in all captions and are also present in the vocabulary of a pretrained version of the NIC model which is used in our experiments.

For all titles, punctuation, stop words and non-ASCII characters are removed, and all characters are converted to lower case. Including all words that occur four or more times in all titles, 14091 unique words remain in the title vocabulary. Table 4.2 reports the top-ten words, excluding stop words ("the", "and", and "of"), and their corresponding counts. The word "sd" stands for South Dakota in this context, an example of the use of "sd" is: "House with Porch at 303 North Bridge Street, Canton SD, Lincoln County".

4.5 Neural Image Caption Model

Since we treat automatic image caption generation as translation, we use a LSTM language model that computes its initial state given an image. See Section 4.5 for an explanation of the LSTM. Specifically, Google's implementation of the NIC model was used, which can be found in the Tensorflow models repository[2]. During training, we fed either an image I or a compressed representation of image and title features as initial input to the LSTM, followed by a word-for-word feeding of the image's description $S_0, ..., S_{n-1}$ (see Figure 4.5).

Fang et al. [65] state that the direct use of captions in training provides the following advantages. Captions contain information that is inherently salient, contain a variety of words, allow the model to learn common sense knowledge about scenes (e.g., a person is more likely to sit on a chair than to stand on it), and the joint multi-modal representation allows for selecting the most suitable description for an image.

A special start <S> and end symbol </S> were appended to the beginning and end of the true caption S, respectively. The input image was converted to features using a CNN. In this case, we used the Inception-V3 architecture [67]. Both the image and the textual input were mapped to the same space using the CNN and a word embedding W_e. The word embedding W_e was learned or retrained during training, depending on whether a pretrained model was used. The NIC model tries

[2]github.com/tensorflow/models/tree/master/research/im2txt

Word	Count
engine	35231
studio	25037
richards	25013
county	24979
number	18054
type	17522
sd	12656
locomotive	8999
house	8757
train	8446

Table 4.2: Top-ten words and corresponding counts in title vocabulary.

to minimize the following cost function:

$$L(I, S) = -\sum_{t=1}^{N} \log p_t(S_t) \tag{4.1}$$

CNNs require a large amount of data for training, and therefore, we use a InceptionV3 network pretrained on the ILSVRC-2012-CLS image classification dataset as released by Google as part of their TF library[3]. Also, we use a pretrained NIC network that has been trained over three million iterations[4]. This NIC network has been trained on the MSCOCO dataset, which comprises over 200,000 labeled image-caption pairs of a wide variety of objects categories[5].

Pretrained models are better at generalizing and we want to make use of this advantage by deploying a pretrained NIC model. The model is trained over 3M iterations on over 123K labeled image-caption pairs from the MSCOCO dataset. However, this means that the model's current input must be compatible with the pretrained NIC model. To this end, we propose an autoencoder that learns to combine image features and title features into a compressed representation. The autoencoder tries to keep this compressed representation as similar to the original image features as possible. This architecture is discussed in Section 4.6.

The Long Short-Term Memory Network

LSTM networks, as introduced by Hochreiter and Schmidhuber [41], do not have a fundamentally different architecture from RNN, however, they can learn long-term dependencies without having the gradients to vanish or explode like with standard RNNs. LSTMs use logical gates to control the flow of information that goes through. Specifically, they determine whether to forget the current cell state value (forget gate f), whether to read its input (input gate i), and whether to output the updated cell state (output gate o). The equations for the gates, cell state and output are defined as follows:

$$f_t = \sigma(W_f \dot{[}h_{t-1}, x_t] + b_f) \tag{4.2}$$

$$i_t = \sigma(W_i \dot{[}h_{t-1}, x_t] + b_i) \tag{4.3}$$

$$o_t = \sigma(W_o \dot{[}h_{t-1}, x_t] + b_0) \tag{4.4}$$

$$c_t = f_t \bullet c_{t-1} + i_t \bullet h(W_c \dot{[}h_{t-1}, x_t] + b_c) \tag{4.5}$$

$$m_t = o_t \bullet c_t \tag{4.6}$$

$$p_{t+1} = softmax(m_t), \tag{4.7}$$

where $\bullet$ represents a product with a gate value and σ and h represent the sigmoid and hyperbolic tangent nonlinearities.

[3] github.com/tensorflow/tensorflow/tree/master/tensorflow/contrib/slim/python/slim/nets

[4] drive.google.com/file/d/0B_qCJ40uBfjEWVItOTdyNUFOMzg/view

[5] cocodataset.org

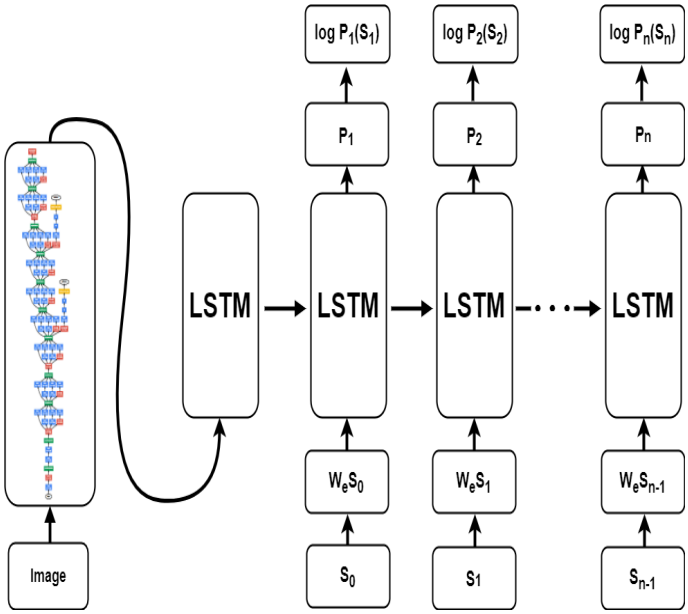


Figure 4.5: NIC model as introduced by Vinyals et al. [6, 10]. The encoder is a convolutional neural network that expects an image. The CNN features are used as initial input for the LSTM encoder.

4.6 Autoencoder Architecture

To incorporate title information into the NIC model, there are several approaches that can be considered. To increase generalization, it is interesting to investigate how to incorporate such textual information in such a way that it can be used by a pretrained NIC model, rather than training the NIC model from scratch. Though, this requires the input data to match the expected input data of the pretrained model. Hence, we propose an autoencoder which learns a compressed representation of the image features and title. An autoencoder is an unsupervised learning algorithm that tries to reconstruct its inputs from a compressed representation in the hidden layer (See Figure 4.6). Essentially, autoencoders reduce input so that only relevant parts are preserved, which is shown by the fact that the input can be reconstructed from the compressed representation. Similar to the NIC model, autoencoders consist of an encoder and a decoder part.

In the image caption scenario, we want to feed a combination of image and title features. To let the combined features vector match the expected NIC input, we set the dimensionality of the original image features as the dimensionality of the hidden layer. Moreover, we want the model to learn compressed representations that match the original image feature as closely as possible. For that reason, we set the weights between the image features input and the hidden layer to an identity matrix $\mathbb{I}$, which has the property that:

$$A\mathbb{I} = A. \tag{4.8}$$

We want to minimize the following combined weighted Mean Squared Error (MSE) loss function:

$$L(\mathbf{I}, \mathbf{T}) = \alpha \frac{1}{N} \sum_{i=1}^{N} (I_i - I_i')^2 + \beta \frac{1}{N} \sum_{i=1}^{N} (T_i - T_i')^2, \tag{4.9}$$

where N is the number of data points, I_i and T_i are the image and title features, and I_i' and T_i' their compressed and reconstructed input, respectively, and α and β are constants that allow each term

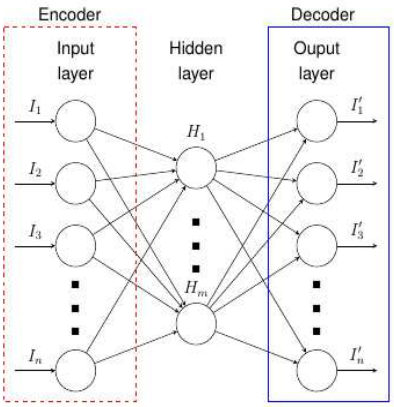


Figure 4.6: General autoencoder architecture.

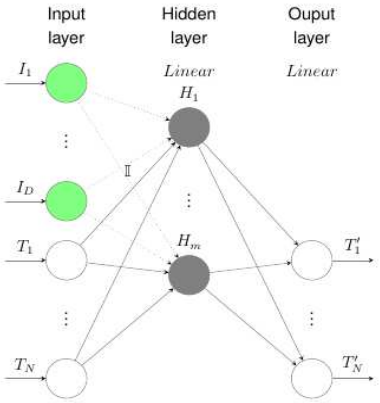


Figure 4.7: Autoencoder that compresses image and title features.

to be given a certain weight. Both the hidden and the output layer use linear activation functions. Figure 4.6 shows a schematic representation of this architecture.

Chapter 5

Experiments and Evaluation

The data used in this work consists of images and high-level captions that describe "meta-information" about the image. This "meta-information" is not directly visible in the image itself. Therefore, one characteristic of the $\text{CONTENT}_{\text{dm}}$ dataset is that images and their captions complement each other. This can be checked by computing the NIC model's perplexity, which is described in Section 5.1. When the image and caption are complementary, they are not necessarily directly related and we expect that training the NIC model from scratch on $\text{CONTENT}_{\text{dm}}$ data yields poor results. Section 5.2 describes the results of training the model from scratch.

Given the data's complementary nature and for better generalization, it is desirable to use a pretrained model. Here, we explore to what extent we can use higher-level title information as input for a pretrained NIC model. To this end, we propose an autoencoder that learns to compress CNN image features and title embedding features, such that they can be fed to a pretrained NIC model. Section 5.3 discusses the results of training the autoencoder. For all further experiments we used smaller subsets (65K and 37K) of the data, because the time it took to generate the `TFRecord` files grew significantly with the amount of data they had to store. As a next experiment, we fine-tune the autoencoder's weights between the title input features and the hidden layer. We reason that updating the model's title encoder weights on the negative log likelihood loss as defined in Equation 4.1 may improve the captions that the model produces. The results of this finetuning step are discussed in Section 5.4. Third, we trained the complete model and evaluate the captions the model generates both with and without the use of title information. The captions that result from a model that is pretrained on more general image-captions pairs, and retrained on $\text{CONTENT}_{\text{dm}}$ data, may include both lower-level and higher-level information about the images. Section 5.5 discusses the results of this procedure. Finally, in Section 5.6, we elaborate on captions as generated by these different model architectures and illustrate this with examples. Evaluation of the natural language output of the model is a fundamentally difficult task [68]. In this work, we let human raters evaluate the image captions on accuracy (i.e., to what extent the caption described what is in the image).

Importantly, for the evaluation we select images with a title of 10 or more words, as this filters out most of the less directly informative titles, such as identifiers like "C00540" in PRPL and many repeated titles like "Richards Studio A110732-1" in TPL.

5.1 Perplexity

Inspection of the data showed that many captions describe higher-level information and do not foucs on the image itself. This suggests that the captions and images are complementary. As mentioned in

Section 4.5, the NIC model tries to minimize the average negative log probability of the target words (See Equation 4.1). Typically, the measure reported in most work is perplexity, which is defined as

$$e^{-\sum_{t=1}^{N} \log p_t(S_t)}. \tag{5.1}$$

Hence, this perplexity measure is the average per-word perplexity which gives the entropy of the distribution. In other words, a perplexity of 12 can be seen as if the model had to choose uniformly and independently among 12 possible words. The perplexity on a validation set of the MSCOCO dataset, on which the pretrained NIC model is trained, is 8.39. The perplexity for a subset of the $\text{CONTENT}_{\text{dm}}$ data is approximately four orders of magnitude higher. This indicates a high level of uncertainty at each time step in the language model to predict a caption given an image from this dataset and suggests that the $\text{CONTENT}_{\text{dm}}$ images and captions are complementary.

5.2 Neural Image Caption on $\text{CONTENT}_{\text{dm}}$ Data

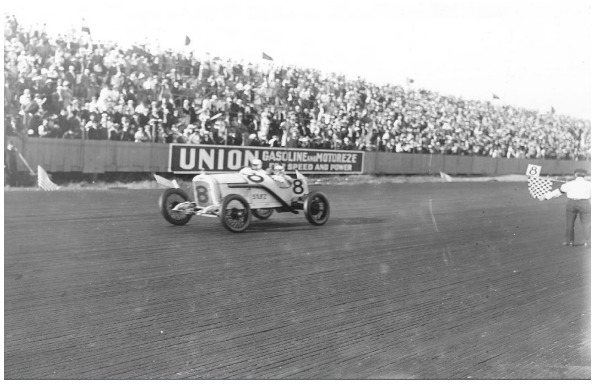
Firstly, we trained the NIC model from scratch on over 200K samples of $\text{CONTENT}_{\text{dm}}$ data (we refer to this model as NIC_{dm}). This was done on a NVIDIA Titan-X GPU and 10K train steps took approximately two hours, we let the model train for a maximum of 600K steps. Although training a model from scratch on $\text{CONTENT}_{\text{dm}}$ data means that the model is less likely to generalize well, it is interesting to see what kind of captions the model learns to generate. We expected the model to highly overfit on the train data, and a quick inspection of the caption generated by the model at 2K, 5K and 10K, and 600K steps confirms this. Namely, it generates captions on test data which repeat training captions and it does not create novel captions. Figure 5.1 shows examples of captions as generated by the NIC_{dm} model after 5K steps. It becomes apparent that the model trained from scratch assigns similar captions to very different images. Although the captions appear to include more details and are extensive, they are just repeated training captions and inaccurate for the corresponding image. Consequently, we leave this model out of the caption evaluation in Section 5.6.

5.3 Autoencoder: Combining Objectives

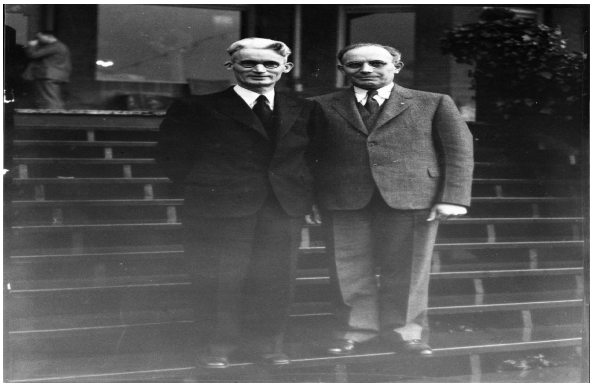
The use of title information as input to the NIC model may contribute to increasing the probability that the model outputs certain words. Almost a third of the words in the title vocabulary also exist in the vocabulary of the pretrained NIC model. By adding the title information as input, it may help increase the probability that the model chooses these words or synonyms. Similar to how the image features contribute by specifying objects and relations between objects, the title can also help steer the model's caption generation towards specific words. To illustrate, in case the model thinks an image depicts a building and the title contains the word "restaurant", which is part of the pretrained model's vocabulary, the probability for this word may increase accordingly.

To that end, we tested whether a compressed representation of the image and title can be learned using an autoencoder. The hidden and output layer are linearly activated and an initial learning rate $\alpha = 0.001$ is used. The learning rate should be initialized relatively small, because we found that the model otherwise keeps overshooting. We set the image weight $\alpha = 10$, and the title weight $\beta = 1$, so that the model emphasizes on retaining the image features whilst adding title information. The results show that the autoencoder is able to learn a compressed representation that closely resembles the original Inception-V3 image features. The model optimized the loss function as shown in

Captions:
1) this is a 5x4 glass plate of two individuals standing in front of a twostory brick building with
2) this is a 5x4 glass plate of two individuals standing in front of a twostory building with a
3) this is a 5x4 glass plate of two individuals standing in front of a singlestory building with a



Captions:
1) this is a 5x4 glass plate of two individuals standing in front of a singlestory building with a
2) this is a 5x4 glass plate of two individuals standing in front of a singlestory building with a
3) this is a 5x4 glass plate of two individuals standing in front of a twostory building with a



Captions:
1) portrait view of mr. <UNK>from the odom photography studio on bradentons old main street .
2) portrait view of mr. <UNK>from the odom photography studio on bradentons old street .
3) portrait view of mr. <UNK>from the odom photography studio on bradentons main street .

Captions:
1) portrait view of mr. <UNK>from the odom photography studio on bradentons old main street .
2) portrait view of ms. <UNK>from the odom photography studio on bradentons old main street .
3) portrait view of mr. <UNK>.

Figure 5.1: Examples of captions as generated by the NIC_{dm} model

Equation 4.9. Figure 5.2 and Figure 5.3 show that the model learns both an compressed representation that closely resembles the original Inception-V3 image features and a reconstruction of the title features.

5.4 Fine-tuning Autoencoder

Next, we combined the autoencoder, as described in the previous section, and the pretrained NIC model. We refer to the model as NIC_{AE}. Training NIC_{AE} for an additional 10K steps took us ap-

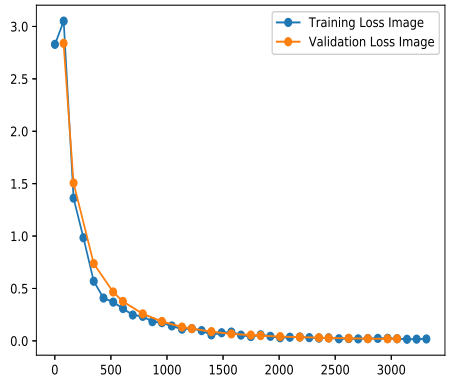


Figure 5.2: The autoencoder's image loss.

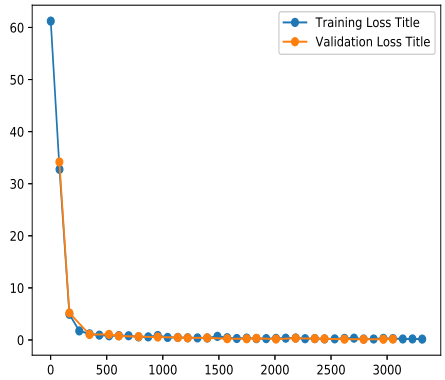


Figure 5.3: The autoencoder's title loss.

proximately 1.5 hours on a NVIDIA Titan-X GPU. The weights between the title input and the hidden layer in the autoencoder were updated by gradient descent optimization on the loss as defined in Equation 4.1. All other parameters were kept fixed. The model's image loss as computed on the train data slightly increased, whereas the model's title loss on the train data slightly decreased. However, a further inspection of the generated captions is needed to indicate how this fine-tuning contributes to the caption generation process.

5.5 Fine-tuning Both Autoencoder and NIC

Then, we also trained both the autoencoder and the NIC model. We refer to this model as NICAE for convenience. For the caption evaluation in Section 5.6, we used a model trained for around 9K steps (NICAE_{9K}) and a model trained for around 81K steps (NICAE_{81K}). The retraining was done with a initial learning rate of $\alpha = 0.03$, because the model did not start from scratch and a larger learning rate would diminish the learning.

5.6 Caption Evaluation

The steps described in the previous sections were performed for the same reason: we want to investigate how the changes we implemented influence the model's final caption generation. Therefore, the most important part of the analysis was to evaluate the generated captions. Given the fact that evaluation of natural language is difficult, we performed a qualitative analysis in which we picked some examples of images and their generated caption and anlayzed them. Additionally, to provide an indication of how accurate captions were given an image, we asked several human raters to evaluate a set of captions given the images.

5.6.1 Baseline

First, we performed an analysis of captions that were generated using the NIC model that was pretrained on MSCOCO data for 3M steps. Below, a handful of randomly selected captions for images

from the $\text{CONTENT}_{\text{dm}}$ dataset are reported and discussed. For all generated captions, we use beam size of 3 unless explicitly stated otherwise. For clarity, we divide the images into three different categories that we came across: 1) people and animals, 2) landscape and street scenes, 3) text and art. Then, we compare the baseline model with the baseline model that uses title input (referred to as baseline_{title}).

People and Animals

A `person` is a category in the MSCOCO dataset. Despite some mistakes, the baseline model is generally good at detecting people and animals in an image (see Figure 5.4). The model does not always distinguish all people in an image correctly. For instance, sometimes the model describes only one age category or only one gender (e.g., see upper right image in Figure 5.4).

Landscape and street scenes

Many images in the $\text{CONTENT}_{\text{dm}}$ dataset are photographs of flyers with text or art pieces. This is something the NIC model is not trained on, and is generally hard to describe both for computers and people. The abstraction level of art makes it difficult to describe it in natural language. An image that contains only text can be described as “this is a piece of text” or “this is a sign”. However, it requires understanding of the text itself, if we want to generate more meaningful captions for images of text. We illustrate the difficulties with some examples in Figure 5.5.

Text and Art

Many images in the $\text{CONTENT}_{\text{dm}}$ dataset are photographs of flyers with text or art pieces. This is something the NIC model is not trained on, and is generally hard to describe both for computers and people. The abstraction level of art makes it difficult to describe it in natural language. An image that contains only text can be described as “this is a piece of text” or “this is a sign”. However, it requires understanding of the text itself, if we want to generate more meaningful captions for images of text. We illustrate the difficulties with some examples in Figure 5.6.

Mistakes

Despite the model’s fair ability to create general captions for images in the $\text{CONTENT}_{\text{dm}}$ dataset, the model also makes mistakes. These mistakes include but are not limited to the following two types: 1) wrongly classified object/action, 2) hallucination. Examples of incorrectly classified objects are shown in Figure 5.7, and examples of hallucination are given in Figure 5.8. The model seems to make up animals (e.g., cows or giraffes) in pictures of a grass field or similar. Moreover, it sometimes misclassifies objects and actions, such as for the left image in Example 5.7, where a book is mistaken for a cell phone and the action of holding the book is mistaken for talking on the cell phone. Generally, manual inspection of the images and their created captions reveals certain objects and actions that occur often (both correctly and incorrectly) in the captions (e.g., people, trains, sitting, and standing).

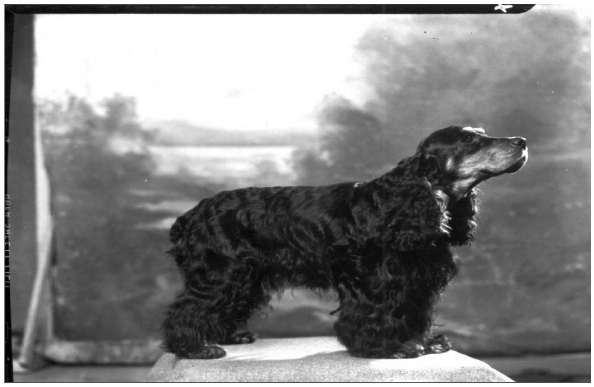
Comparison of Baseline and Baseline with Title

We manually analyzed captions for eight images for the baseline model and the $\text{baseline}_{\text{title}}$ model. If we strictly compared the captions in the order in which they are generated, 14 out of 24 captions were different for the two models. However, if we compared the captions without taking into account

Captions:
1) a young boy wearing a tie and a shirt .
2) a young boy is sitting on a couch .
3) a young boy wearing a tie and a shirt

Captions:
1) a group of children sitting on a bench .
2) a black and white photo of a group of children .
3) a black and white photo of a group of children

Captions:
1) a black and white photo of people on motorcycles .
2) a black and white photo of people on motorcycles
3) a black and white photo of a group of people .

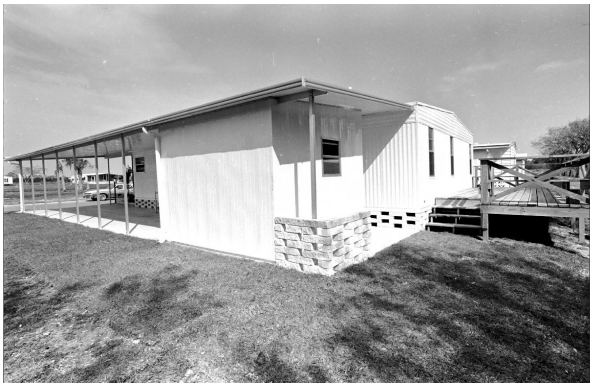


Captions:
1) a black and white dog standing on top of a field .
2) a black and white dog standing on top of a dirt field .
3) a black and white dog standing on top of a grass covered field .

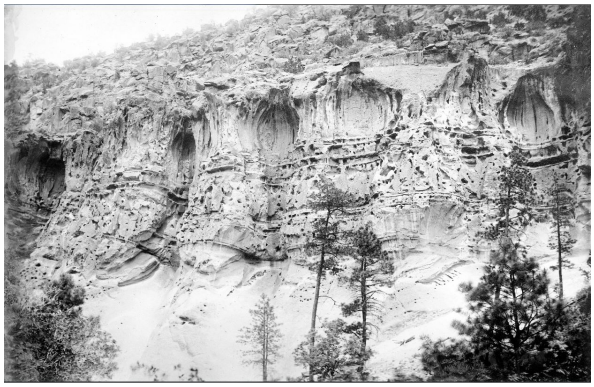
Figure 5.4: People and animals baseline examples.

the order only seven out of 24 captions are different. Table 5.1 reports these differences. From these examples, we did not see any significant changes in terms of novel captions being generated when using the title as compared to the baseline. The differences seem to indicate that the small change in image features results in similar captions with minor changes, such as slightly different word order or the use of different, yet, related words.

Captions:
1) a black and white photo of a city street .
2) a black and white photo of a city street
3) a black and white photo of a street corner



Captions:
1) a black and white photo of a house
2) a black and white photo of a barn and a barn
3) a black and white photo of a barn and a house



Captions:
1) a black and white photo of a snowy landscape .
2) a black and white photo of a snowy mountain .
3) a black and white photo of a snowy landscape

Captions:
1) a black and white photo of a train on a track .
2) a black and white photo of a train on a track
3) a black and white photo of a street with a mountain in the background .

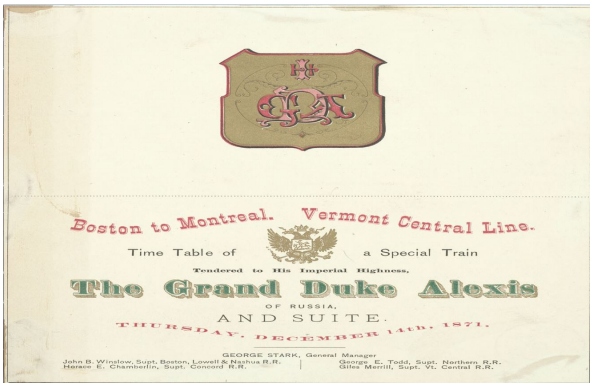
Figure 5.5: Landscape and street baseline examples.

5.6.2 Evaluation of NIC_{AE}

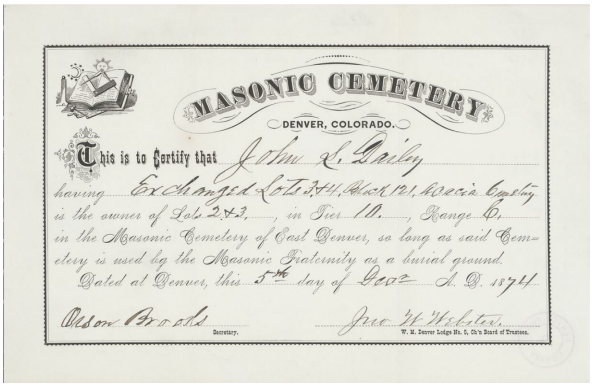
In this section, we explore some captions generated by the NIC_{AE} model. First, we found some instances where mistakes are corrected. For example, the mistakenly detected cell phone in the left image of Figure 5.7 is corrected. The new captions completely leave out the cell phone: 1) a woman sitting on a bench in front of a building, 2) a woman sitting on a bench in front of a brick wall, 3) a woman sitting on a bench in front of a store.

5.6.3 Evaluation of NICAE

The captions resulting from the $NICAE_{81K}$ do not seem to be an improvement over the $NICAE_{9K}$ model's captions. The $NICAE_{81K}$ model both with and without title mostly creates captions that seem to be a result of the model's overfitting on the train data, and are less accurate. For instance, "this is a <UNK >slide of the front facade of a toilet building ." and "interior view of the <UNK>of the



Captions:
1) a sign that says " <UNK><UNK>" on a table .
2) a sign that says " <UNK><UNK>" on the side of a building .
3) a sign that says " <UNK><UNK>" is on a table .



Captions:
1) a black and white photo of a book on a table
2) a black and white photo of a book on a table .
3) a black and white photo of a book shelf with books

Captions:
1) a statue of a bear sitting on a bench .
2) a statue of a bear sitting on a bench
a statue of a bear sitting on a wooden bench .

Captions:
1) a painting of a man on a skateboard .
2) a painting of a man on a surfboard .
3) a painting of a man on a skateboard

Figure 5.6: Text and art baseline examples.

<UNK>of the <UNK>.". However, we found one exception which showed an interesting result in the caption created by the NICAE$_{81K}$ model when using titles. In this case, the model outputs a detail that is correct given the image, that is not found in the captions of the other models. Figure 5.9 shows the image and the caption as generated by the NICAE$_{81K}$ both with and without use of title for comparison. Although the model with title includes the "U.S. air force" detail, which is correct here, this information is not directly found in the title. In the rest of this section, we elaborate on the captions created by the NICAE$_{9K}$ model.

An example of where the NICAE$_{9K}$ model with title provides a more detailed caption is shown in Figure 5.10. Another interesting example where the NICAE9K model produced a caption where the title is directly reflected is shown in Figure 5.11. Additionally, there are examples to be found that show a slightly more extensive sentence (see Figure 5.13). The NICAE$_{9K}$ model produced a caption that specifies the location of the clock more specifically (e.g., "on the front of the building") as compared to the baseline (e.g., with a clock tower"). There are more examples where the NICAE$_{9K}$ model produces more detailed captions as compared to the baseline. Figure 5.14 shows that the NICAE$_{9K}$

Captions:
1) a woman sitting on a bench talking on a cell phone .
2) a woman sitting on a bench talking on a phone .
3) a woman sitting on a bench talking on the phone .

Captions:
1) a young boy sitting on a chair with a teddy bear .
2) a young boy sitting on a couch with a teddy bear .
3) a young boy sitting on a couch with a stuffed animal .

Figure 5.7: Examples (underlined) of incorrect object and action classification in baseline examples.

model describes not only the people, but also the building behind them, whereas the baseline only describes the people.

There are, however, examples that show that the model overfits to the train data and that the captions become less accurate. For instance, many captions start with “This is a photograph of...” or “This is a slide of a”, which are repeated parts of train captions. There is a lot of repetition within the captions as well. Captions often look like the following: “a <UNK>of the <UNK>of the <UNK>of the <UNK>.” and “portrait view of the front end of the front of the front view of the front of the”. Also, the content of the captions becomes incorrect, such as the example in Figure 5.15.

	Baseline	**Baseline$_{\text{title}}$**
1	a black and white photo of a large field	a black and white photo of a train yard
2	a black and white photo of a group of people on a mountain .	a black and white photo of people on a beach
3	a black and white photo of a train on a track .	a black and white photo of a street with a mountain range .
4	a black and white photo of a train on a track	a black and white photo of a road with a mountain in the background .
5	a young boy is sitting on a couch .	a young boy wearing a tie and a sweater .
6	a young boy wearing a tie and a shirt	a young boy is wearing a tie and a shirt

Table 5.1: Examples of differences in captions between baseline and baseline$_{\text{title}}$, with the difference underlined.

Captions:
1) a black and white photo of cows grazing in a field .
2) a black and white photo of cows grazing in a pasture .
3) a black and white photo of cows grazing in a field

Captions:
1) a giraffe is standing in a grassy field .
2) a giraffe is standing in the middle of a field .
3) a giraffe is standing in a grassy field

Captions:
1) a red fire hydrant sitting in front of a building .
2) a red fire hydrant sitting in front of a brick building .
3) a red fire hydrant sitting in the middle of a garden .

Captions:
1) a black and white photo of a train on a track .
2) a black and white photo of a train on the tracks .
3) a black and white photo of a train on a track

Figure 5.8: Examples of hallucination in baseline examples.

5.7 Evaluation by Human Raters

To get an indication of how accurate the captions are, we asked human raters to judge captions via a questionnaire (See Appendix A). However, even with several different people rating captions, the evaluation remains highly subjective and prone to people's own interpretation. Therefore, this evaluation is used to gain more insight into the image caption's quality, but does not lend itself for hard conclusions.

Design Choices

The evaluation of captions can be a tedious task, especially when several models create several captions and the number of comparisons grows. Therefore, we decided to include only one caption per model per image, instead of all three captions. We want to compare six different model initializations:

Figure 5.9: Caption $NICAE_{81K}$ without title: an airplane is parked on the runway .
Caption $NICAE_{81K}$ with title: an aerial view of the u.s. air force .
Title: A helicopter at Manatee Memorial Hospital to transfer a patient to Tampa General Hospital

Figure 5.10: Caption baseline: a black and white photo of a city street .
Caption $NICAE_{9K}$ without title: this is a photograph of a street in the city .
Caption $NICAE_{9K}$ with title: this is a photograph of a street sign on the side of the building .
Title: Looking west along south side of Manatee Avenue from 9th St. W. to Professional Building.

- NIC_ae;
- NICAE_9K no title;
- NICAE_9K title;
- NICAE_81K no titlel
- NICAE_81K title;
- Baseline.

Moreover, the length of the questionnaire required participants to focus for a longer time (approximately 30 minutes) and compare many different captions can be a strenuous task. To make sure

Figure 5.11: Caption baseline: a black and white photo of a man and a woman
Caption NICAE$_{9K}$ without title: this is a portrait of a man and a woman pose in front of a building .
Caption NICAE$_{9K}$ with title: this is a portrait of a man and a woman pose in front of a porch .
Title: A woman and man standing; he is holding baby seated on porch rail

Figure 5.12: Caption baseline: a black and white photo of a park bench .
Caption NICAE$_{9K}$ with title: this is a view of a bridge in the distance .

participants read the questionnaire items carefully, we included a control question that asked participants to perform a simple calculation. Additionally, the evaluation of image captions is highly subjective in nature. In this questionnaire we only focused on one dimension (e.g., how accurate is a caption for an image). In this case, accuracy was defined as *"An image caption's accuracy refers to the extent with which the caption describes what is depicted in the image. It is NOT about grammatical correctness"*. At the end of the questionnaire, participants were asked again how an image's accuracy was defined. Finally, the scale for one item in the questionnaire was incorrectly implemented as 5-points Likert scale instead of 7-point Likert scale. Therefore, those scores were corrected with the following calculation:

$$x_{new} = (B - A) * (x_{old} - a)/(b - a) + A, \tag{5.2}$$

where A and B are the minimum and maximum of the old scale (1 and 5, respectively), a and b are

Figure 5.13: Caption baseline: a black and white photo of a church with a clock tower .
Caption NICAE_{9K} with title: this is a photograph of a tall building with a clock on the front of the building .
Title: Looking west along south side of Manatee Avenue from 9th St. W. to Professional Building.

Figure 5.14: Caption baseline: a group of people standing next to each other .
Caption NICAE_{9K} with title: a group of men and women standing in front of a large building with a flag on the

the minimum and maximum of the new scale (1 and 7, respectively), and x is the score to convert.

Participants

12 participants (7 male and 5 female) took part in an online questionnaire in which they rated how accurate captions were given an image. The participants were recruited via convenience sampling and were between 23 and 62 years of age (mean = 31.58, SD = 13.06). The participants were asked to rate 300 captions (50 images, 6 captions each) on their accuracy on a 7-points Likert scale (1 = completely inaccurate, 7 = completely accurate).

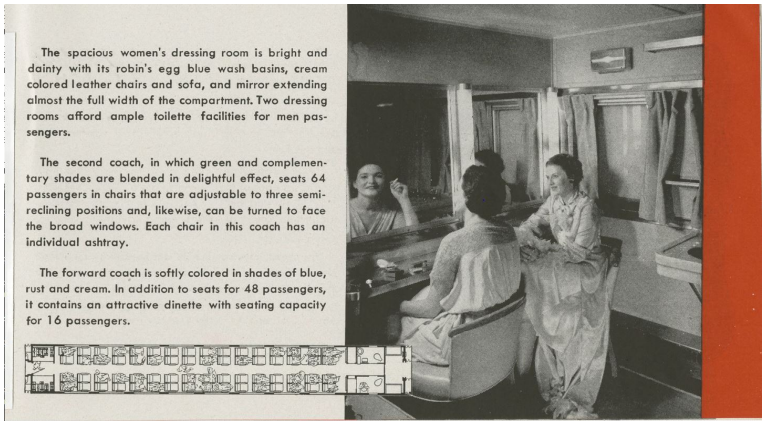


Figure 5.15: Caption baseline: a group of people standing in front of a building .
Caption NICAE_{9K} with title: this is a photograph of the president of the united states of the <UNK>.
Title: Page 9

Results

All participants answered the control calculation question correctly, confirming that they did not went through the questionnaire without reading the items. Only one out of 12 participants did not indicate the right definition of an image's accuracy. However, Krippendorff's alpha α was computed to indicate inter-rater reliablity and $\alpha = 0.68$, indicating that human raters vary in their ratings. The results show that the baseline model and the NICAE_{9K} without title model perform best in terms of how accurate they describe the CONTENT_{dm} images (see Table 5.2). Overall, the scores were low, which indicates that for this dataset there is little consensus on the captions' accuracy.

Model	Mean	Standard Deviation
NIC_ae	2.80	1.80
NICAE_9K no title	3.25	2.14
NICAE_9K title	2.44	1.60
NICAE_81K no title	2.36	1.45
NICAE_81K title	2.32	1.44
Baseline	3.33	1.90

Table 5.2: The mean and standard deviation for the accuracy scores from the questionnaire.

Chapter 6

Discussion and Future Work

In this work, we have focused on automatic image caption generation for OCLC's CONTENT_{dm} dataset. In line with the approach of Vinyals et al. [6, 10], we have approached this task as a translation problem, in which an LSTM language model generates captions given an image. The novelty of this work is in the use of title information, which is combined with the image as initial input to the LSTM model. However, there are some limitations to this study which are discussed below.

It can be argued that this work has tried to tackle two different problems in one. Namely, exploring automatic image caption generation on a novel dataset (the CONTENT_{dm} dataset) and exploring the feasibility of including title information in a pretrained state-of-the-art image captioning model. This work provides proof-of-concept that the title information can be combined with image features and be used as initial input to a language model. This provides an answer to the first research question, of how title information can be used in a pretrained NIC model for OCLC's CONTENT_{dm} images. However, the second research question on what effect the use of title information has on the caption generation, has proven difficult to answers. The current results are mixed and hard conclusions cannot be drawn, however, it does provide many suggestions for future research.

The CONTENT_{dm} dataset can be further refined according to the titles and captions. In this work, the collections have been selected on their ostensible similarity, however, they have not been further refined. In future work, it would be useful to start off with a more specific selection of images. Although, refinements may lead to smaller datasets, it can be argued that a dataset's refinement outweighs its size, as long as a reasonable amount of images remain in the refined dataset. Refinement could focus on creating a dataset that still contains higher-level captions, but has a certain trade-off between variety between images and enough similarity between image-captions pairs for the model to learn a pattern. On the other hand, it could be beneficial to further investigate and filter out entries that do not contribute any information. In other words, it would be useful to analyze the set of images with captions (or titles) that are repeated often, and see if those textual description are still informative or whether they are mostly noise.

Moreover, we used OCLC's CONTENT_{dm} Word-2-Vec (W2V) embedding model on the titles. There are two directions to include in further studies. Namely, we averaged the embeddings over all words in a title, and it could be interesting to explore other ways to treat the title. In line with further refining the dataset, for instance, important keywords could be extracted from titles and words that are less likely to contribute high-level information, could be filtered out. For this, it would be interesting to do a further analysis of the words, which ones co-occur most often, explore synonyms, and so on. The captions embeddings, however, were learned on a different dataset (the MSCOCO dataset). It would be interesting to study the use of embedding models, such W2V, for the captions as well. These embedding models take the context of words into account and may therefore increase

the model's ability to use synonyms that are close in the semantic space. Additionally, we retrained the NIC model using the pretrained vocabulary size. In future work, this may be extended to also include the words that are specific for the $\text{CONTENT}_{\text{dm}}$ dataset rather than disregarding them. In combination with a better refined dataset, this may lead to interesting, novel captions.

Furthermore, the evaluation of the output of a natural language generation model like NIC is a difficult task. However, it would be interesting and necessary to further analyze the captions that the model generates. In this work, we performed a qualitative, explorative analysis of the captions, but in the future it is necessary to look further into inexpensive metrics to quantitize the results. There are several evaluation measures for this, such as the BLEU score [69] and METEOR score [70], however, they require a set of reference captions per image. Moreover, the questionnaire that was used focused on one dimension of a caption's quality: how accurate is the caption in terms of giving a description of what is in the image. It is important to note that this is not the only aspect that contributes to a caption's quality, and for instance, future analyses should also include other dimensions, such a grammatical correctness or the correctness of the "meta-information" in a caption, based on the application the caption will be used for. Especially, when we come to a point where captions do include "meta-information" about an image, it is important to perform evaluation with experts that have the knowledge to accurately rate the information's correctness. Nevertheless, the model has not yet properly learned the pattern in this data and there is still a large discrepancy between the true and predicted captions. Therefore, it is important to first investigate whether we can improve the caption generation quality on this type of data. As stated above, a starting point would be to further refine the dataset before using it in training.

Additionally, title information is only one of the options to add to the model, and it is interesting to explore the use of different types of information, such as the date, location, or other "meta-information" that is available for the image. The $\text{CONTENT}_{\text{dm}}$ database has numerous entries that contain such information. Hence, it would help gain insight into what information can be used to improve the automatic image caption generation process, and how this information should be treated.

Chapter 7

Conclusion

There has been a considerable growth of the volume of digital collections, which has lead to an increasing demand for automated techniques that support the management, navigation and search of these collections. As machine learning techniques are advancing, it becomes feasible to automatically generate image captions. However, one of the main challenges that needs to be addressed is to create captions that include higher-level information, such as the event or location shown in an image. Recently, the combination of a convolutional neural network and a recurrent language model has shown to achieve state-of-the-art on image captioning. However, one of the main challenges that needs to be addressed is to create captions that include higher-level information, such as the event or location shown in an image.

In this work, we assumed that images and text naturally co-occur and explored the feasibility of including title information in a pretrained state-of-the-art image captioning model using OCLC's $\text{CONTENT}_{\text{dm}}$ data. An autoencoder learned to generate a compressed representation of image and title features, which is compatible as input to a pretrained state-of-the-art image captioning model. Evaluation of the output of such a natural language generation model is difficult, however, we found some interesting examples. On the one hand, for instance, in some cases the model's captions were more detailed if the title information was used in the input. On the other hand, many captions still suffered from the model's overfitting to the training data and were mostly repeated training captions. Hence, the results are still mixed and an evaluation by 12 human raters showed that there was little consensus on the captions' accuracy. Despite that results are preliminary, the exploration in this thesis provides interesting insights into the automatic generation of higher-level image captions and a basis for future work.

Bibliography

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

[2] L. Fei-Fei, A. Iyer, C. Koch, and P. Perona, "What do we perceive in a glance of a real-world scene?" *Journal of vision*, vol. 7, no. 1, pp. 10–10, 2007.

[3] Y. Feng and M. Lapata, "Visual information in semantic representation," in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 2010, pp. 91–99.

[4] J. Fan, Y. Gao, and H. Luo, "Hierarchical classification for automatic image annotation," in *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2007, pp. 111–118.

[5] B. Z. Yao, X. Yang, L. Lin, M. W. Lee, and S.-C. Zhu, "I2t: Image parsing to text description," *Proceedings of the IEEE*, vol. 98, no. 8, pp. 1485–1508, 2010.

[6] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3156–3164.

[7] P. Héde, P.-A. Moëllic, J. Bourgeoys, M. Joint, and C. Thomas, "Automatic generation of natural language descriptions for images," in *Coupling approaches, coupling media and coupling languages for information retrieval*. LE CENTRE DE HAUTES ETUDES INTERNATIONALES D'INFORMATIQUE DOCUMENTAIRE, 2004, pp. 306–313.

[8] Y. Feng and M. Lapata, "Automatic caption generation for news images," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 4, pp. 797–812, 2013.

[9] ——, "How many words is a picture worth? automatic caption generation for news images," in *Proceedings of the 48th annual meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2010, pp. 1239–1249.

[10] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: Lessons learned from the 2015 mscoco image captioning challenge," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 4, pp. 652–663, 2017.

[11] R. Bernardi, R. Cakici, D. Elliott, A. Erdem, E. Erdem, N. Ikizler-Cinbis, F. Keller, A. Muscat, and B. Plank, "Automatic description generation from images: A survey of models, datasets, and evaluation measures." *J. Artif. Intell. Res.(JAIR)*, vol. 55, pp. 409–442, 2016.

[12] P. Kuznetsova, V. Ordonez, T. Berg, and Y. Choi, "Treetalk: Composition and compression of trees for image descriptions," *Transactions of the Association of Computational Linguistics*, vol. 2, no. 1, pp. 351–362, 2014.

[13] B. Cheng and D. M. Titterington, "Neural networks: A review from a statistical perspective," *Statistical science*, pp. 2–30, 1994.

[14] J. E. Dayhoff and J. M. DeLeo, "Artificial neural networks," *Cancer*, vol. 91, no. S8, pp. 1615–1635, 2001.

[15] W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *The bulletin of mathematical biophysics*, vol. 5, no. 4, pp. 115–133, 1943.

[16] L. Medsker and L. C. Jain, *Recurrent neural networks: design and applications*. CRC press, 1999.

[17] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural networks*, vol. 2, no. 5, pp. 359–366, 1989.

[18] C. M. Bishop, "Pattern recognition and machine learning (information science and statistics) springer-verlag new york," *Inc. Secaucus, NJ, USA*, 2006.

[19] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[20] X. Zhu, "Semi-supervised learning," in *Encyclopedia of machine learning*. Springer, 2011, pp. 892–897.

[21] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *nature*, vol. 323, no. 6088, p. 533, 1986.

[22] J. F. Kolen and J. B. Pollack, "Back propagation is sensitive to initial conditions," in *Advances in neural information processing systems*, 1991, pp. 860–867.

[23] A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," in *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, 2014, pp. 1764–1772.

[24] S. Venugopalan, H. Xu, J. Donahue, M. Rohrbach, R. Mooney, and K. Saenko, "Translating videos to natural language using deep recurrent neural networks," *arXiv preprint arXiv:1412.4729*, 2014.

[25] T. Mikolov, M. Karafiát, L. Burget, J. Cernockỳ, and S. Khudanpur, "Recurrent neural network based language model." in *Interspeech*, vol. 2, 2010, p. 3.

[26] A. Graves *et al.*, *Supervised sequence labelling with recurrent neural networks*. Springer, 2012, vol. 385.

[27] A. Karpathy, J. Johnson, and L. Fei-Fei, "Visualizing and understanding recurrent networks," *arXiv preprint arXiv:1506.02078*, 2015.

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

[29] K. Irie, Z. Tüske, T. Alkhouli, R. Schlüter, and H. Ney, "Lstm, gru, highway and a bit of attention: An empirical overview for language modeling in speech recognition." in *INTERSPEECH*, 2016, pp. 3519–3523.

[30] A. Graves and J. Schmidhuber, "Offline handwriting recognition with multidimensional recurrent neural networks," in *Advances in neural information processing systems*, 2009, pp. 545–552.

[31] A. Graves, "Generating sequences with recurrent neural networks," *arXiv preprint arXiv:1308.0850*, 2013.

[32] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in neural information processing systems*, 2014, pp. 3104–3112.

[33] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[34] R. Wang, M. Panju, and M. Gohari, "Classification-based rnn machine translation using grus," *arXiv preprint arXiv:1703.07841*, 2017.

[35] T. Mikolov and G. Zweig, "Context dependent recurrent neural network language model." *SLT*, vol. 12, pp. 234–239, 2012.

[36] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3128–3137.

[37] M. Sundermeyer, R. Schlüter, and H. Ney, "Lstm neural networks for language modeling," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[38] P. J. Werbos, "Backpropagation through time: what it does and how to do it," *Proceedings of the IEEE*, vol. 78, no. 10, pp. 1550–1560, 1990.

[39] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE transactions on neural networks*, vol. 5, no. 2, pp. 157–166, 1994.

[40] R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training recurrent neural networks," in *International Conference on Machine Learning*, 2013, pp. 1310–1318.

[41] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[42] S. Lawrence, C. L. Giles, A. C. Tsoi, and A. D. Back, "Face recognition: A convolutional neural-network approach," *IEEE transactions on neural networks*, vol. 8, no. 1, pp. 98–113, 1997.

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

[44] G. Levi and T. Hassner, "Age and gender classification using convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2015, pp. 34–42.

[45] M. Thoma, "Analysis and optimization of convolutional neural network architectures," *arXiv preprint arXiv:1707.09725*, 2017.

[46] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[47] Y. Yang, C. L. Teo, H. Daumé III, and Y. Aloimonos, "Corpus-guided sentence generation of natural images," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2011, pp. 444–454.

[48] T. Uricchio, L. Ballan, L. Seidenari, and A. Del Bimbo, "Automatic image annotation via label transfer in the semantic space," *Pattern Recognition*, vol. 71, pp. 144–157, 2017.

[49] X.-J. Wang, L. Zhang, X. Li, and W.-Y. Ma, "Annotating images by mining image search results," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 11, pp. 1919–1932, 2008.

[50] J. Johnson, L. Ballan, and F.-F. Li, "Love thy neighbors: Image annotation by exploiting image metadata," *arXiv preprint arXiv:1508.07647*, 2015.

[51] G. Singh and J. Kaur, "Survey of image object recognition techniques," 2017.

[52] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[53] T.-H. Chan, K. Jia, S. Gao, J. Lu, Z. Zeng, and Y. Ma, "Pcanet: A simple deep learning baseline for image classification?" *IEEE Transactions on Image Processing*, vol. 24, no. 12, pp. 5017–5032, 2015.

[54] Y. LeCun, K. Kavukcuoglu, and C. Farabet, "Convolutional networks and applications in vision," in *Circuits and Systems (ISCAS), Proceedings of 2010 IEEE International Symposium on*. IEEE, 2010, pp. 253–256.

[55] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.

[56] M. Lin, Q. Chen, and S. Yan, "Network in network," *arXiv preprint arXiv:1312.4400*, 2013.

[57] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[58] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[59] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning." in *AAAI*, 2017, pp. 4278–4284.

[60] L. Ferres, A. Parush, S. Roberts, and G. Lindgaard, “Helping people with visual impairments gain access to graphical information through natural language: The igraph system,” in *ICCHP*. Springer, 2006, pp. 1122–1130.

[61] K. Ramnath, S. Baker, L. Vanderwende, M. El-Saban, S. N. Sinha, A. Kannan, N. Hassan, M. Galley, Y. Yang, D. Ramanan *et al.*, “Autocaption: Automatic caption generation for personal photos,” in *Applications of Computer Vision (WACV), 2014 IEEE Winter Conference on*. IEEE, 2014, pp. 1050–1057.

[62] A. Kojima, T. Tamura, and K. Fukunaga, “Natural language description of human activities from video images based on concept hierarchy of actions,” *International Journal of Computer Vision*, vol. 50, no. 2, pp. 171–184, 2002.

[63] 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 *European conference on computer vision*. Springer, 2010, pp. 15–29.

[64] S. Li, G. Kuflkarni, T. L. Berg, A. C. Berg, and Y. Choi, “Composing simple image descriptions using web-scale n-grams,” in *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*. Association for Computational Linguistics, 2011, pp. 220–228.

[65] H. Fang, S. Gupta, F. Iandola, R. Srivastava, L. Deng, P. Dollár, J. Gao, X. He, M. Mitchell, J. Platt *et al.*, “From captions to visual concepts and back,” 2015.

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

[67] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the inception architecture for computer vision,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2818–2826.

[68] E. Reiter and A. Belz, “An investigation into the validity of some metrics for automatically evaluating natural language generation systems,” *Computational Linguistics*, vol. 35, no. 4, pp. 529–558, 2009.

[69] 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*. Association for Computational Linguistics, 2002, pp. 311–318.

[70] 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*, 2005, pp. 65–72.

Appendix A

Title of first appendix

Rating of Image Captions

This form is used to evaluate image captions that are automatically generated.

In this questionnaire, we ask you to rate how accurate six different captions are for an image.

NOTE: An image caption's accuracy refers to the extent with which the caption describes what is depicted in the image. It is NOT about grammatical correctness.

Please, answer as quickly as possible without overthinking your answer. There are no right or wrong answers.

The questionnaire takes approximately 30 minutes to complete.

Thank you for your participation!

* Required

Untitled Section

General information

For research purposes we would like to know your gender and age.

1. **Gender** *
   Mark only one oval.
   - ◯ Female
   - ◯ Male
   - ◯ Prefer not to say
   - ◯ Other: _____

2. **Age** *
   _____

Image 1


EXIT
4 NORTH

3. *

Please indicate how accurate you think the following captions describe the image on a scale from 1 - completely inaccurate to 7 - completely accurate.
Mark only one oval per row.

	1 - Completely inaccurate	2 - Mostly inaccurate	3 - Slightly inaccurate	4 - Neither inaccurate nor accurate	5 - Slightly accurate	6 - Mostly accurate	7 - Completely accurate
a group of people sitting around a table .							
a man and a woman pose in front of a table .							
a group of men and women standing in front of a building .							
a man and woman pose in front of a table in a room .							
a group of men and women pose in front of a building .							
a group of people sitting in a room .							

4. **Comments**

If you have any comments about this question, you can leave them here.

Image 2

5. *

Please indicate how accurate you think the following captions describe the image on a scale from 1 - completely inaccurate to 7 - completely accurate.
Mark only one oval per row.

	1 - Completely inaccurate	2 - Mostly inaccurate	3 - Slightly inaccurate	4 - Neither inaccurate nor accurate	5 - Slightly accurate	6 - Mostly accurate	7 - Completely accurate
a park bench sitting next to a river .							
this is a view of a bridge in the distance .							
this is a view of a bridge in the distance .							
view of a bridge over the river , to .							
this is a <UNK> photograph of a bridge over a river . that by near , l .							
a black and white photo of a park bench .							

6. **Comments**

If you have any comments about this question, you can leave them here.

Image 3

7. *

Please indicate how accurate you think the following captions describe the image on a scale from 1 - completely inaccurate to 7 - completely accurate.
Mark only one oval per row.

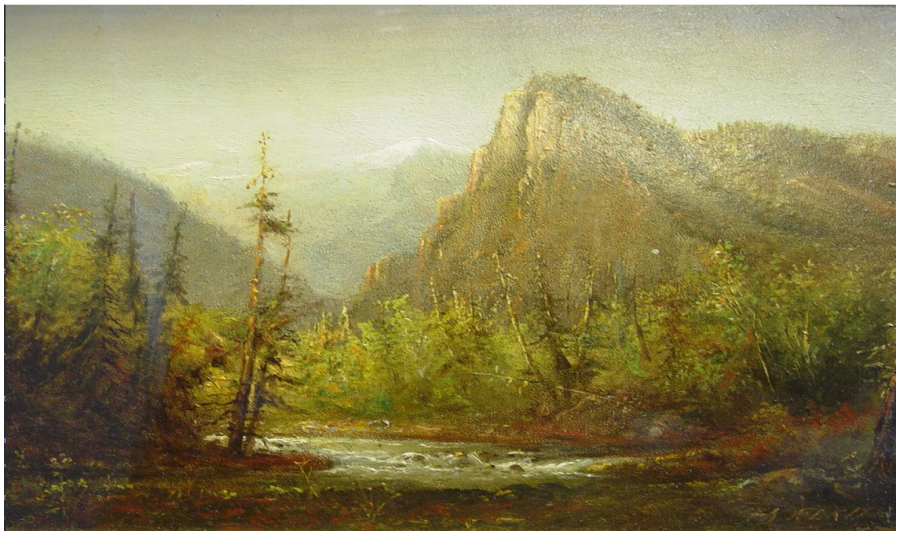
	1 - Completely inaccurate	2 - Mostly inaccurate	3 - Slightly inaccurate	4 - Neither inaccurate nor accurate	5 - Slightly accurate	6 - Mostly accurate	7 - Completely accurate
a woman sitting on a bench in front of a building .							
this is a woman sitting on a bench in front of a brick building .							
this is a portrait of a man and a woman sitting on a bench .							
this is a <UNK> photograph of a man and a woman pose in front of a window .							
this is a <UNK> photograph of a woman sitting on a bench .							
a woman sitting on a bench talking on a cell phone .							

8.

Comments

If you have any comments about this question, you can leave them here.

Image 4



9. *

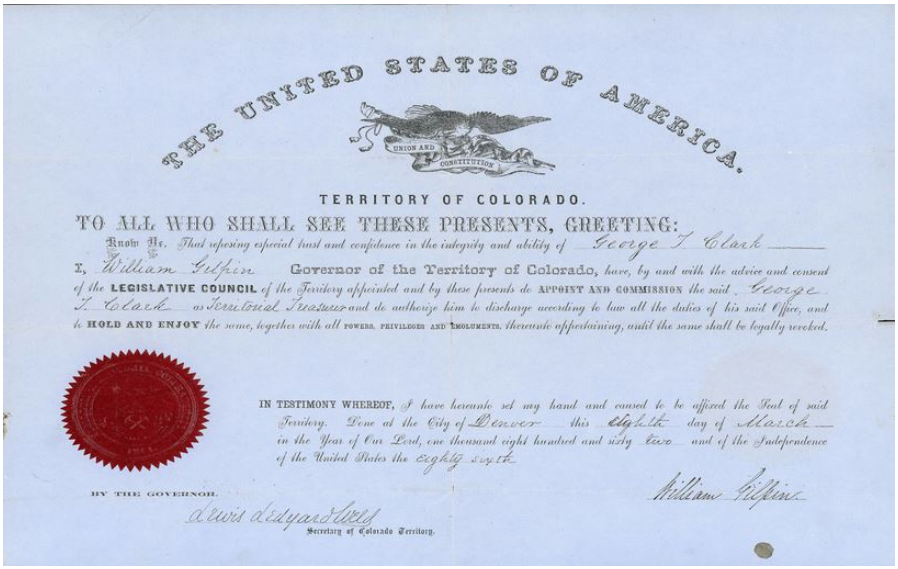
Please indicate how accurate you think the following captions describe the image on a scale from 1 - completely inaccurate to 7 - completely accurate.
Mark only one oval per row.

	1 - Completely inaccurate	2 - Mostly inaccurate	3 - Slightly inaccurate	4 - Neither inaccurate nor accurate	5 - Slightly accurate	6 - Mostly accurate	7 - Completely accurate
a group of people sitting on a bench near a river .							
this is a woman sitting on a bench in front of a brick building .							
this is a view of a river in the distance .							
this is a <UNK> photograph of a bridge over a river . that by near , l .							
this is a <UNK> photograph of a bridge in the distance . that by near , l .							
a view of a river with a bridge in the background .							

10. **Comments**

If you have any comments about this question, you can leave them here.

Image 5



THE UNITED STATES OF AMERICA.

UNION AND CONSTITUTION

TERRITORY OF COLORADO.

TO ALL WHO SHALL SEE THESE PRESENTS, GREETING:

Know Ye, That reposing especial trust and confidence in the integrity and ability of George T. Clark

I, William Gilpin Governor of the Territory of Colorado, have, by and with the advice and consent of the LEGISLATIVE COUNCIL of the Territory appointed and by these presents do APPOINT AND COMMISSION the said George T. Clark as Territorial Treasurer and do authorize him to discharge according to law all the duties of his said Office, and to HOLD AND ENJOY the same, together with all POWERS, PRIVILEGES AND EMOLUMENTS thereunto appertaining, until the same shall be legally revoked.

IN TESTIMONY WHEREOF, I have hereunto set my hand and caused to be affixed the Seal of said Territory. Done at the City of Denver this Eighth day of March in the Year of Our Lord, one thousand eight hundred and sixty five and of the Independence of the United States the Eighty sixth

BY THE GOVERNOR.

Lewis Ledyard Weld
Secretary of Colorado Territory.

William Gilpin

11. *

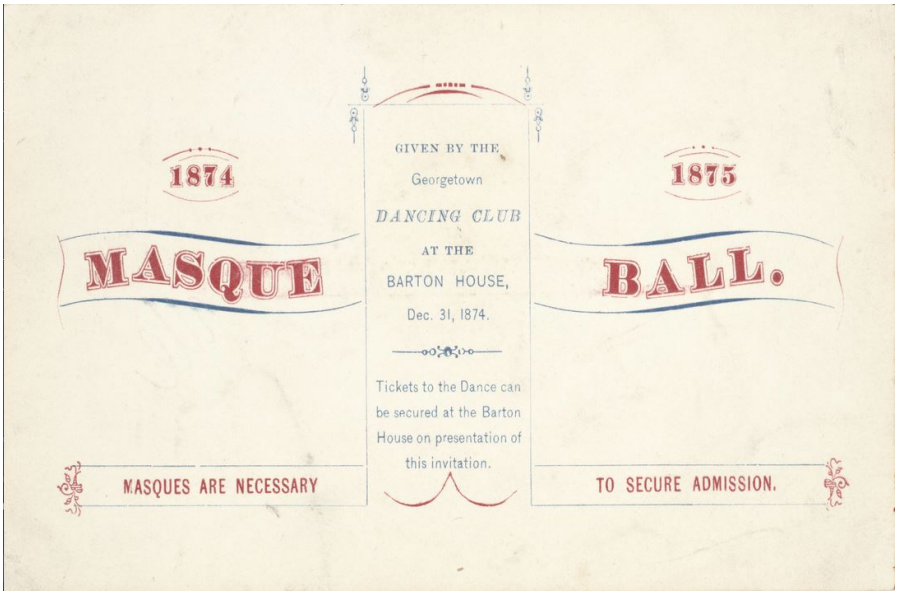
Please indicate how accurate you think the following captions describe the image on a scale from 1 - completely inaccurate to 7 - completely accurate.
Mark only one oval per row.

	1 - Completely inaccurate	2 - Mostly inaccurate	3 - Slightly inaccurate	4 - Neither inaccurate nor accurate	5 - Slightly accurate	6 - Mostly accurate	7 - Completely accurate
a close up of a person holding a cell phone							
this is a slide of a laptop computer on a table .							
this is a slide of a slide of a building .							
this is a <UNK> photograph of a man in a suit and tie .							
this is a slide of the front of a laptop computer .							
a book sitting on top of a wooden table .							

12. **Comments**

If you have any comments about this question, you can leave them here.

Image 6



13. *

Please indicate how accurate you think the following captions describe the image on a scale from 1 - completely inaccurate to 7 - completely accurate.
Mark only one oval per row.

	1 - Completely inaccurate	2 - Mostly inaccurate	3 - Slightly inaccurate	4 - Neither inaccurate nor accurate	5 - Slightly accurate	6 - Mostly accurate	7 - Completely accurate
a collage of photos with a sign on it	⬭	⬭	⬭	⬭	⬭	⬭	⬭
this is a slide of the front end of a toilet .	⬭	⬭	⬭	⬭	⬭	⬭	⬭
this is a slide of the front end of a <UNK> .	⬭	⬭	⬭	⬭	⬭	⬭	⬭
this is a <UNK> slide of the front of the building . that by near , I .	⬭	⬭	⬭	⬭	⬭	⬭	⬭
this is a slide of the front of the front facade of a toilet .	⬭	⬭	⬭	⬭	⬭	⬭	⬭
a close up of a book on a table	⬭	⬭	⬭	⬭	⬭	⬭	⬭

14. **Comments**

If you have any comments about this question, you can leave them here.

Image 7

15. *

Please indicate how accurate you think the following captions describe the image on a scale from 1 - completely inaccurate to 7 - completely accurate.
Mark only one oval per row.

	1 - Completely inaccurate	2 - Mostly inaccurate	3 - Slightly inaccurate	4 - Neither inaccurate nor accurate	5 - Slightly accurate	6 - Mostly accurate	7 - Completely accurate
a group of people standing around a table .							
a group of men and women standing in front of a train .							
a group of men and women standing in front of a building .							
a group of men and women standing in front of a building .							
a group of men and women standing in front of a building .							
a group of people standing in a kitchen .							

16.

Comments

If you have any comments about this question, you can leave them here.

Image 8


WARDS
Reinhardt's
MUSIC
BOOTS
WORK
CLOTHING
RECORDS
NO LEFT TURN

17. *

Please indicate how accurate you think the following captions describe the image on a scale from 1 - completely inaccurate to 7 - completely accurate.
Mark only one oval per row.

	1 - Completely inaccurate	2 - Mostly inaccurate	3 - Slightly inaccurate	4 - Neither inaccurate nor accurate	5 - Slightly accurate	6 - Mostly accurate	7 - Completely accurate
a black and white photo of a street sign .							
this is a photograph of a street in the city .							
this is a photograph of a street sign on the side of the building .							
this is a pole slide of a tall building with a sign on the side of the building							
this is a <UNK> photograph of a street sign on the side of the street .							
a black and white photo of a city street .							

18.

Comments

If you have any comments about this question, you can leave them here.

Image 9

19. *

Please indicate how accurate you think the following captions describe the image on a scale from 1 - completely inaccurate to 7 - completely accurate.
Mark only one oval per row.

	1 - Completely inaccurate	2 - Mostly inaccurate	3 - Slightly inaccurate	4 - Neither inaccurate nor accurate	5 - Slightly accurate	6 - Mostly accurate	7 - Completely accurate
a black and white photo of a city street .							
this is a photograph of the front end of a train .							
this is a photograph of the front entrance of the building .							
this is a <UNK> photograph of the front facade of a toilet brick building . that by parking							
black and white photograph of the front end of the front of the building .							
a black and white photo of a train station .							

20.

Comments

If you have any comments about this question, you can leave them here.

Image 10

21. *

Please indicate how accurate you think the following captions describe the image on a scale from 1 - completely inaccurate to 7 - completely accurate.
Mark only one oval per row.

	1 - Completely inaccurate	2 - Mostly inaccurate	3 - Slightly inaccurate	4 - Neither inaccurate nor accurate	5 - Slightly accurate	6 - Mostly accurate	7 - Completely accurate
a group of people walking down a street .							
this is a photograph of a person in the distance .							
this is a photograph of the <UNK> of the building in the background .							
this is a <UNK> photograph of a man and a woman in front of a tall building with							
this is a <UNK> photograph of a tall building with a sign on the side of the building							
a black and white photo of a city street .							

22.

Comments
If you have any comments about this question, you can leave them here.

Image 11

23. *

Please indicate how accurate you think the following captions describe the image on a scale from 1 - completely inaccurate to 7 - completely accurate.
Mark only one oval per row.

	1 - Completely inaccurate	2 - Mostly inaccurate	3 - Slightly inaccurate	4 - Neither inaccurate nor accurate	5 - Slightly accurate	6 - Mostly accurate	7 - Completely accurate
a group of people sitting around a wooden bench .							
this is a portrait of a man and a woman pose in front of a brick building with							
this is a portrait of a man and a woman pose in front of a tall building with							
this is a <UNK> photograph of a man in a suit and tie . he wears a suit							
this is a portrait of a man and a woman pose in front of a brick building with							
a group of men standing next to each other .							

24. **Comments**

If you have any comments about this question, you can leave them here.

Image 12

25. *

Please indicate how accurate you think the following captions describe the image on a scale from 1 - completely inaccurate to 7 - completely accurate. Your answers are important to us. To make sure you are still paying attention we ask you to perform a simple calculation below.
Mark only one oval per row.

	1 - Completely inaccurate	2 - Mostly inaccurate	3 - Slightly inaccurate	4 - Neither inaccurate nor accurate	5 - Slightly accurate	6 - Mostly accurate	7 - Completely accurate
a black and white photo of men in a field .							
a group of men with a field in the distance .							
a group of men standing next to each other .							
a group of men standing in a field							
please enter the answer to 2+2 in the comments							
we want to make sure you are still paying attention							

26. **Comments**

If you have any comments about this question, you can leave them here.

Image 13

27. *

Please indicate how accurate you think the following captions describe the image on a scale from 1 - completely inaccurate to 7 - completely accurate.
Mark only one oval per row.

	1 - Completely inaccurate	2 - Mostly inaccurate	3 - Slightly inaccurate	4 - Neither inaccurate nor accurate	5 - Slightly accurate	6 - Mostly accurate	7 - Completely accurate
a table that has a bunch of items on it							
this is a slide of a slide of a train .							
this is a slide of a slide of a building with a toilet and a sign on the							
this is a <UNK> photograph of a tall building with a covered front porch .							
this is a slide of the front end of a train . that by near , I .							
a red stop sign sitting in the middle of a field .							

28. **Comments**

If you have any comments about this question, you can leave them here.

Image 14

29. *

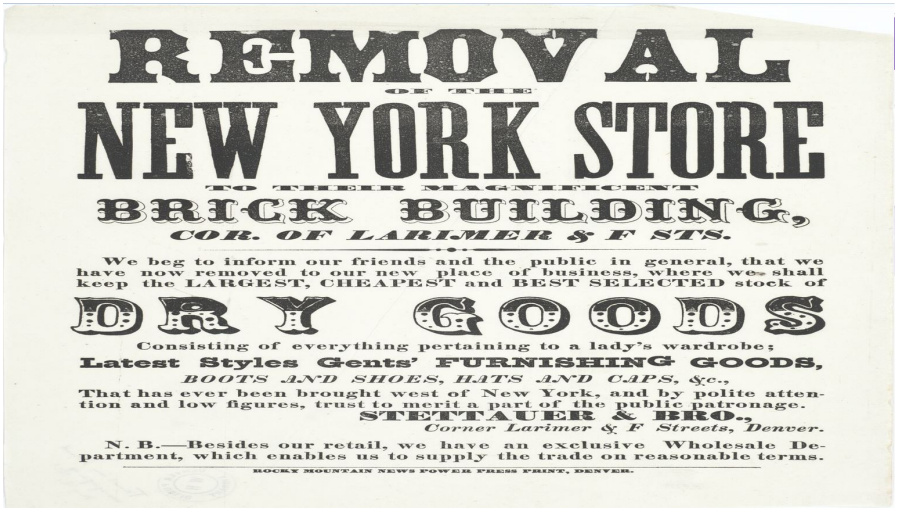
Please indicate how accurate you think the following captions describe the image on a scale from 1 - completely inaccurate to 7 - completely accurate.
Mark only one oval per row.

	1 - Completely inaccurate	2 - Mostly inaccurate	3 - Slightly inaccurate	4 - Neither inaccurate nor accurate	5 - Slightly accurate	6 - Mostly accurate	7 - Completely accurate
a bunch of street signs on a pole							
this is a photograph of a street sign in front of a tree .							
this is a photograph of a person in the background .							
this is a <UNK> photograph of a single story frame building with a covered front porch .							
this is a <UNK> photograph of an individual that by near , l .							
a close up of a street sign with a sky background							

30. **Comments**

If you have any comments about this question, you can leave them here.

Image 15



REMOVAL
OF THE
NEW YORK STORE
TO THEIR MAGNIFICENT
BRICK BUILDING,
COR. OF LARIMER & F STS.

We beg to inform our friends and the public in general, that we have now removed to our new place of business, where we shall keep the LARGEST, CHEAPEST and BEST SELECTED stock of

DRY GOODS

Consisting of everything pertaining to a lady's wardrobe;
Latest Styles Gents' FURNISHING GOODS,
BOOTS AND SHOES, HATS AND CAPS, &c.,
That has ever been brought west of New York, and by polite attention and low figures, trust to merit a part of the public patronage.
STETTAUER & BRO.,
Corner Larimer & F Streets, Denver.

N. B.—Besides our retail, we have an exclusive Wholesale Department, which enables us to supply the trade on reasonable terms.

ROCKY MOUNTAIN NEWS POWER PRESS PRINT, DENVER.

31. *

Please indicate how accurate you think the following captions describe the image on a scale from 1 - completely inaccurate to 5 - completely accurate.
Mark only one oval per row.

	1 - Completely inaccurate	2 - Mostly inaccurate	3 - Slightly inaccurate	3 - Neither inaccurate nor accurate	5 - Slightly accurate	6 - Mostly accurate	7 - Completely accurate
a man sitting on a bench reading a newspaper .							
this is a photograph of a <UNK> of <UNK> .							
this is a slide of the front of the building .							
this is a <UNK> photograph of a sign on the side of a building . that by near							
this is a photograph of the <UNK> of <UNK> .							
a picture of a book that is on a table .							

32.

Comments

If you have any comments about this question, you can leave them here.

Image 16

33. *

Please indicate how accurate you think the following captions describe the image on a scale from 1 - completely inaccurate to 7 - completely accurate.
Mark only one oval per row.

	1 - Completely inaccurate	2 - Mostly inaccurate	3 - Slightly inaccurate	4 - Neither inaccurate nor accurate	5 - Slightly accurate	6 - Mostly accurate	7 - Completely accurate
a group of people standing in front of a building .							
this is a photo of the front door of the house .							
this is a photograph of a man in a suit and tie .							
this is a <UNK> photograph of a man in a suit and tie in front of a toilet							
this is a <UNK> glass plate of a single story glass plate .							
a black and white photo of a vase of flowers .							

34. **Comments**

If you have any comments about this question, you can leave them here.

Image 17

35. *

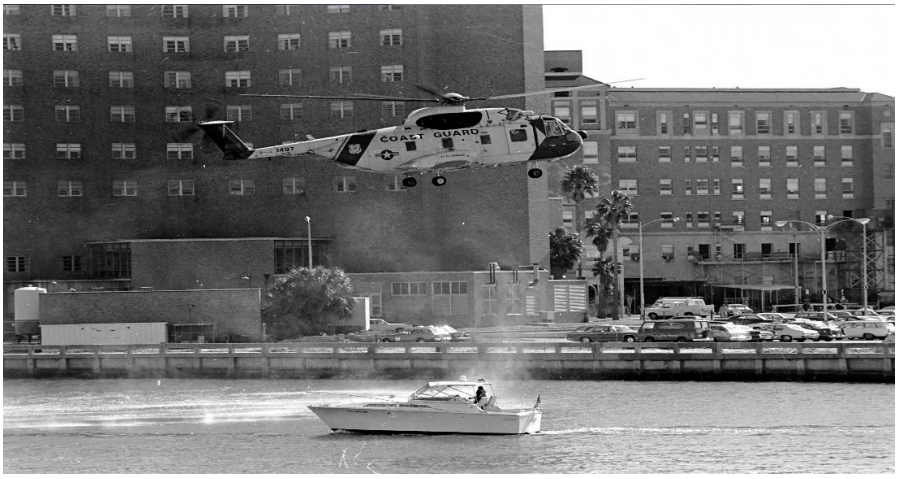
Please indicate how accurate you think the following captions describe the image on a scale from 1 - completely inaccurate to 7 - completely accurate.
Mark only one oval per row.

	1 - Completely inaccurate	2 - Mostly inaccurate	3 - Slightly inaccurate	4 - Neither inaccurate nor accurate	5 - Slightly accurate	6 - Mostly accurate	7 - Completely accurate
a birthday cake with a sign on top of it							
this is a slide of the front end of a toilet .							
this is a slide of the front of the building .							
this is a photograph of the <UNK> of <UNK> .							
this is a slide of the front facade of a toilet building .							
a sign that says `` <UNK> <UNK> '' on a table							

36. **Comments**

If you have any comments about this question, you can leave them here.

Image 18



37. *

Please indicate how accurate you think the following captions describe the image on a scale from 1 - completely inaccurate to 7 - completely accurate.
Mark only one oval per row.

	1 - Completely inaccurate	2 - Mostly inaccurate	3 - Slightly inaccurate	4 - Neither inaccurate nor accurate	5 - Slightly accurate	6 - Mostly accurate	7 - Completely accurate
a group of people standing around a fire hydrant .							
view of an airplane flying over the water .							
view of the front end of a ship .							
this is a <UNK> photograph of an airplane on the ground .							
this is a <UNK> photograph of the <UNK> of the building . that by near , l .							
an airplane flying over a city with a building .							

38. **Comments**

If you have any comments about this question, you can leave them here.

Image 19

39. *

Please indicate how accurate you think the following captions describe the image on a scale from 1 - completely inaccurate to 7 - completely accurate.
Mark only one oval per row.

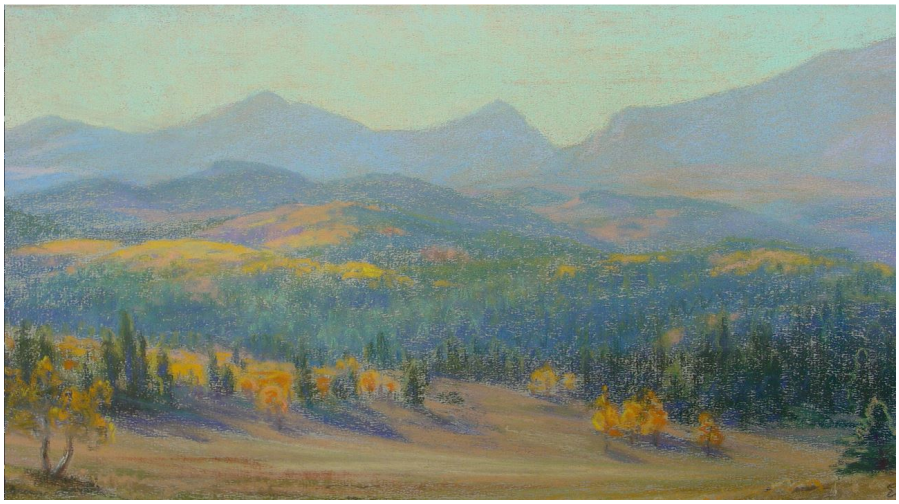
	1 - Completely inaccurate	2 - Mostly inaccurate	3 - Slightly inaccurate	4 - Neither inaccurate nor accurate	5 - Slightly accurate	6 - Mostly accurate	7 - Completely accurate
a man and a woman pose for a picture .							
a man and woman pose for a picture in a park .							
this is a portrait of a man and a woman pose in front of a tree .							
this is a <UNK> photograph of a man and woman pose in front of a house .							
a man and woman pose in front of a tree .							
a man and a woman standing next to each other .							

40.

Comments

If you have any comments about this question, you can leave them here.

Image 20



41. *

Please indicate how accurate you think the following captions describe the image on a scale from 1 - completely inaccurate to 7 - completely accurate.
Mark only one oval per row.

	1 - Completely inaccurate	2 - Mostly inaccurate	3 - Slightly inaccurate	4 - Neither inaccurate nor accurate	5 - Slightly accurate	6 - Mostly accurate	7 - Completely accurate
a group of people standing on top of a hill .							
this is a view of a mountain in the distance .							
this is a slide of the front end of a rocky cliff .							
view of the canyon , to , to .							
this is a <UNK> photograph of a rocky mountain range .							
a couple of horses are standing in a field							

42. **Comments**

If you have any comments about this question, you can leave them here.

Image 21

43. *

Please indicate how accurate you think the following captions describe the image on a scale from 1 - completely inaccurate to 7 - completely accurate.
Mark only one oval per row.

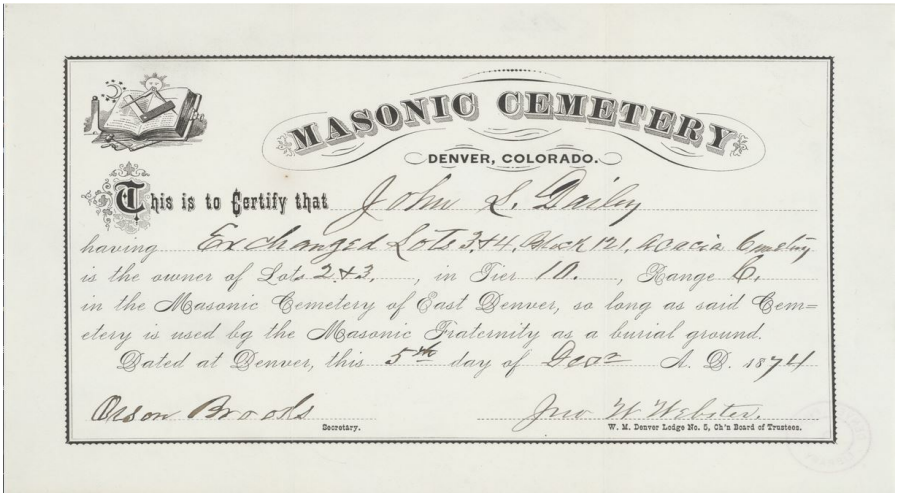
	1 - Completely inaccurate	2 - Mostly inaccurate	3 - Slightly inaccurate	4 - Neither inaccurate nor accurate	5 - Slightly accurate	6 - Mostly accurate	7 - Completely accurate
a black and white photo of a group of people .							
a black and white photograph of a group of men and women .							
a group of men and women pose in front of a house .							
black and white photograph of a man and a woman pose in front of a group of men							
a group of men and women pose in front of a house .							
a black and white photo of a group of people .							

44.

Comments

If you have any comments about this question, you can leave them here.

Image 22



45. *

Please indicate how accurate you think the following captions describe the image on a scale from 1 - completely inaccurate to 7 - completely accurate.
Mark only one oval per row.

	1 - Completely inaccurate	2 - Mostly inaccurate	3 - Slightly inaccurate	4 - Neither inaccurate nor accurate	5 - Slightly accurate	6 - Mostly accurate	7 - Completely accurate
a sign that says `` <UNK> <UNK> " on it .							
this is a photograph of the <UNK> of the <UNK> .							
this is a slide of the front end of a <UNK> .							
this is a slide of the front end of a laptop .							
this is a <UNK> photograph of the <UNK> of <UNK> . that by near , l .							
a black and white photo of a book on a table							

46. **Comments**

If you have any comments about this question, you can leave them here.

Image 23

47. *

Please indicate how accurate you think the following captions describe the image on a scale from 1 - completely inaccurate to 7 - completely accurate.
Mark only one oval per row.

	1 - Completely inaccurate	2 - Mostly inaccurate	3 - Slightly inaccurate	4 - Neither inaccurate nor accurate	5 - Slightly accurate	6 - Mostly accurate	7 - Completely accurate
a large group of people standing on a beach .							
this is a view of a bridge over the water .							
view of a bridge over the river , to .							
aerial view of the city of the river , to .							
aerial view of the bridge , to .							
a black and white photo of a book on a table							

48.

Comments

If you have any comments about this question, you can leave them here.

Image 24

49. *

Please indicate how accurate you think the following captions describe the image on a scale from 1 - completely inaccurate to 7 - completely accurate.
Mark only one oval per row.

	1 - Completely inaccurate	2 - Mostly inaccurate	3 - Slightly inaccurate	4 - Neither inaccurate nor accurate	5 - Slightly accurate	6 - Mostly accurate	7 - Completely accurate
a pair of scissors sitting on top of a table .							
this is a portrait of a man in a suit and tie .							
this is a portrait of a man in a suit and tie .							
this is a <UNK> photograph of an individual in front of a sign .							
this is a slide of the front end of a <UNK> . that by near , I .							
a close up of a pair of scissors on a table							

50. **Comments**

If you have any comments about this question, you can leave them here.

Image 25

51. *

Please indicate how accurate you think the following captions describe the image on a scale from 1 - completely inaccurate to 7 - completely accurate.
Mark only one oval per row.

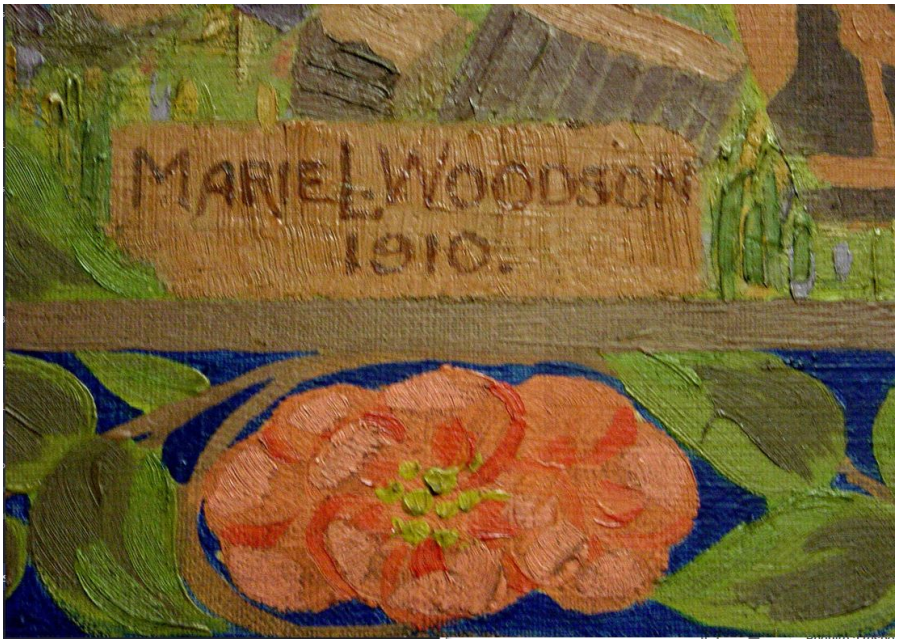
	1 - Completely inaccurate	2 - Mostly inaccurate	3 - Slightly inaccurate	4 - Neither inaccurate nor accurate	5 - Slightly accurate	6 - Mostly accurate	7 - Completely accurate
a group of men standing next to each other .							
two men standing next to each other in a field .							
a group of men and women standing in front of a building .							
this is a <UNK> photograph of a man in a suit and tie . a man stands on							
this is a <UNK> photograph of a man in a suit and tie . a man stands on							
a couple of men standing next to each other .							

52.

Comments

If you have any comments about this question, you can leave them here.

Image 26


MARIE L. WOODSON
1910.

53. *

Please indicate how accurate you think the following captions describe the image on a scale from 1 - completely inaccurate to 7 - completely accurate.
Mark only one oval per row.

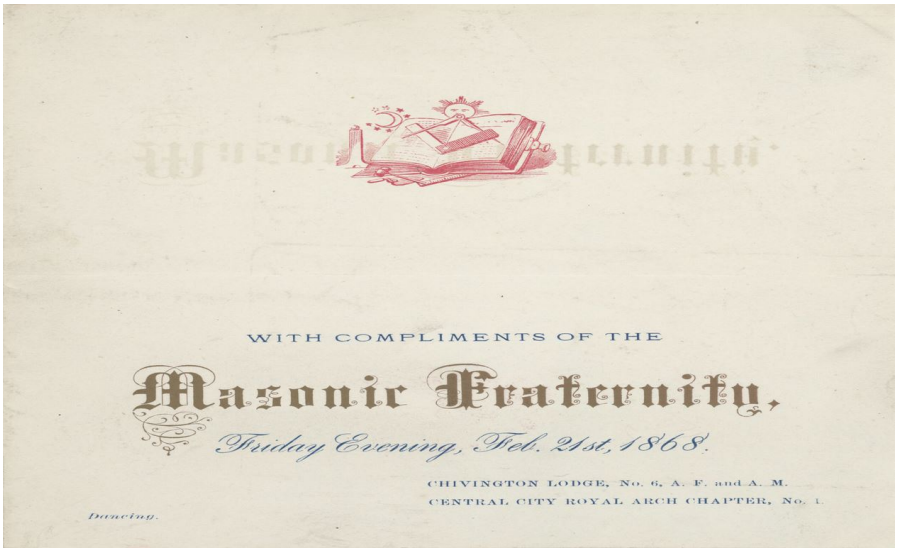
	1 - Completely inaccurate	2 - Mostly inaccurate	3 - Slightly inaccurate	4 - Neither inaccurate nor accurate	5 - Slightly accurate	6 - Mostly accurate	7 - Completely accurate
a sign that is on the side of a building .							
this is a slide of a slide of a toilet building with a wooden roof .							
this is a slide of a slide of a building with a sign on it .							
this is a <UNK> photograph of a man in a suit and tie . that by near ,							
this is a slide of the front facade of a toilet building .							
a statue of a bear sitting on a bench .							

54.

Comments

If you have any comments about this question, you can leave them here.

Image 27



55. *

Please indicate how accurate you think the following captions describe the image on a scale from 1 - completely inaccurate to 7 - completely accurate.
Mark only one oval per row.

	1 - Completely inaccurate	2 - Mostly inaccurate	3 - Slightly inaccurate	4 - Neither inaccurate nor accurate	5 - Slightly accurate	6 - Mostly accurate	7 - Completely accurate
a close up of a sign on a wall							
this is a photograph of the <UNK> of <UNK> .							
this is a slide of the <UNK> of the <UNK> .							
this is a photograph of a man in a suit and tie .							
this is a slide of the front end of the front end of a laptop .							
a close up of a sign on a wall							

56. **Comments**

If you have any comments about this question, you can leave them here.

Image 28

57. *

Please indicate how accurate you think the following captions describe the image on a scale from 1 - completely inaccurate to 7 - completely accurate.
Mark only one oval per row.

	1 - Completely inaccurate	2 - Mostly inaccurate	3 - Slightly inaccurate	4 - Neither inaccurate nor accurate	5 - Slightly accurate	6 - Mostly accurate	7 - Completely accurate
a group of men standing next to each other .							
this is a portrait of a man in a suit and tie .							
<UNK> <UNK> , <UNK> , <UNK> .							
this is a <UNK> photograph of a man in a suit and tie . he wears a suit							
his is a portrait of a man in a suit and tie .							
a painting of a man on a skateboard .							

58.

Comments

If you have any comments about this question, you can leave them here.

Image 29

59. *

Please indicate how accurate you think the following captions describe the image on a scale from 1 - completely inaccurate to 7 - completely accurate.v
Mark only one oval per row.

	1 - Completely inaccurate	2 - Mostly inaccurate	3 - Slightly inaccurate	4 - Neither inaccurate nor accurate	5 - Slightly accurate	6 - Mostly accurate	7 - Completely accurate
a group of people standing around a table .							
a group of men and women standing in front of a table in a room .							
view of the president of the united states of the <UNK> , to .							
a group of men and women stand in front of a store .							
a group of men and women pose in front of a group of men and women standing in							
a group of people standing around a table .							

60. **Comments**

If you have any comments about this question, you can leave them here.

Image 30

61. *

Please indicate how accurate you think the following captions describe the image on a scale from 1 - completely inaccurate to 7 - completely accurate.
Mark only one oval per row.

	1 - Completely inaccurate	2 - Mostly inaccurate	3 - Slightly inaccurate	4 - Neither inaccurate nor accurate	5 - Slightly accurate	6 - Mostly accurate	7 - Completely accurate
a group of men standing next to each other .							
a group of men and women pose for a photograph .							
portrait view of the president of the <UNK> .							
portrait view of <UNK> <UNK> , <UNK> , <UNK> , <UNK> , <UNK> , <UNK> , <UNK> ,							
a group of men and women pose in front of a christmas tree .							
a group of men standing next to each other .							

62.

Comments

If you have any comments about this question, you can leave them here.

Image 31

63. *

Please indicate how accurate you think the following captions describe the image on a scale from 1 - completely inaccurate to 7 - completely accurate.
Mark only one oval per row.

	1 - Completely inaccurate	2 - Mostly inaccurate	3 - Slightly inaccurate	4 - Neither inaccurate nor accurate	5 - Slightly accurate	6 - Mostly accurate	7 - Completely accurate
a group of people standing next to each other .							
black and white photograph of a group of men and women standing in the grass .							
a group of men and women standing in front of a building .							
a group of men and women pose on a dirt road in front of a house .							
a group of men and women pose in front of a group of men and women pose in							
a black and white photo of a group of people .							

64.

Comments

If you have any comments about this question, you can leave them here.

Image 32

65. *

Please indicate how accurate you think the following captions describe the image on a scale from 1 - completely inaccurate to 7 - completely accurate.
Mark only one oval per row.

	1 - Completely inaccurate	2 - Mostly inaccurate	3 - Slightly inaccurate	4 - Neither inaccurate nor accurate	5 - Slightly accurate	6 - Mostly accurate	7 - Completely accurate
a group of people walking down a street .							
this is a photograph of the front entrance of the building .							
this is a photograph of the front entrance of the building .							
this is a pole slide of a tall building with a sign on the side of the building							
black and white photograph of a house in the distance .							
a black and white photo of a town							

66. **Comments**

If you have any comments about this question, you can leave them here.

Image 33

67. *

Please indicate how accurate you think the following captions describe the image on a scale from 1 - completely inaccurate to 7 - completely accurate.
Mark only one oval per row.

	1 - Completely inaccurate	2 - Mostly inaccurate	3 - Slightly inaccurate	4 - Neither inaccurate nor accurate	5 - Slightly accurate	6 - Mostly accurate	7 - Completely accurate
a sign that says `` <UNK> '' on a table .							
this is a photograph of a book on the floor .							
this is a photograph of the president of the united states of the <UNK> .							
this is a photograph of the president of the president of the <UNK> .							
this is a slide of the front end of a toilet . that by near , l .							
a close up of a book on a bed							

68.

Comments

If you have any comments about this question, you can leave them here.

Image 34

69. *

Please indicate how accurate you think the following captions describe the image on a scale from 1 - completely inaccurate to 7 - completely accurate.
Mark only one oval per row.

	1 - Completely inaccurate	2 - Mostly inaccurate	3 - Slightly inaccurate	4 - Neither inaccurate nor accurate	5 - Slightly accurate	6 - Mostly accurate	7 - Completely accurate
a black and white photo of a group of people .							
black and white photograph of men and women standing in front of a building .							
a group of men and women pose in front of a building .							
a group of men and women pose in front of a brick building with a covered porch .							
a group of men and women pose in front of a building .							
a group of men standing next to each other .							

70. **Comments**

If you have any comments about this question, you can leave them here.

Image 35

71. *

Please indicate how accurate you think the following captions describe the image on a scale from 1 - completely inaccurate to 7 - completely accurate.
Mark only one oval per row.

	1 - Completely inaccurate	2 - Mostly inaccurate	3 - Slightly inaccurate	4 - Neither inaccurate nor accurate	5 - Slightly accurate	6 - Mostly accurate	7 - Completely accurate
a group of people standing next to an airplane .							
an airplane is parked on the runway at the airport .							
an airplane parked on the tarmac at the airport .							
an aerial view of the u.s. air force .							
an airplane is parked on the runway .							
a black and white photo of an air plane							

72. **Comments**

If you have any comments about this question, you can leave them here.

Image 36

73. *

Please indicate how accurate you think the following captions describe the image on a scale from 1 - completely inaccurate to 7 - completely accurate.
Mark only one oval per row.

	1 - Completely inaccurate	2 - Mostly inaccurate	3 - Slightly inaccurate	4 - Neither inaccurate nor accurate	5 - Slightly accurate	6 - Mostly accurate	7 - Completely accurate
a group of men standing next to each other .							
two men in suits and ties standing in front of a podium .							
<UNK> , <UNK> , <UNK> , <UNK> , <UNK> , <UNK> , <UNK> , <UNK> ,							
ca .							
president of the president of the president of the united states .							
a group of men standing next to each other .							

74. **Comments**

If you have any comments about this question, you can leave them here.

Image 37

75. *

Please indicate how accurate you think the following captions describe the image on a scale from 1 - completely inaccurate to 7 - completely accurate.
Mark only one oval per row.

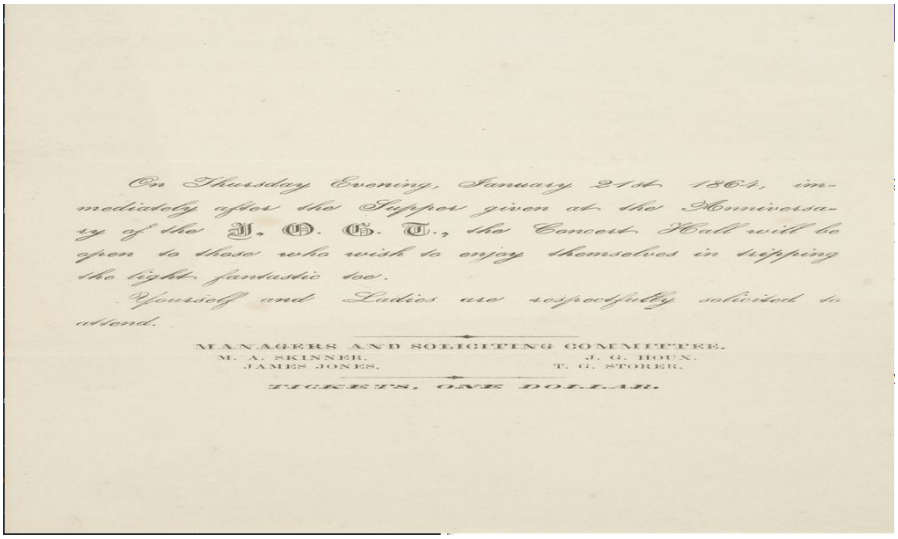
	1 - Completely inaccurate	2 - Mostly inaccurate	3 - Slightly inaccurate	4 - Neither inaccurate nor accurate	5 - Slightly accurate	6 - Mostly accurate	7 - Completely accurate
a group of men standing next to each other .							
a group of men and women pose for a photograph in a room .							
a group of men and women standing in front of a building .							
a man and woman pose in front of a <UNK> .							
a group of men and women pose in front of a group of people in a room .							
a black and white photo of a group of people .							

76.

Comments

If you have any comments about this question, you can leave them here.

Image 38



77. *

Please indicate how accurate you think the following captions describe the image on a scale from 1 - completely inaccurate to 7 - completely accurate.
Mark only one oval per row.

	1 - Completely inaccurate	2 - Mostly inaccurate	3 - Slightly inaccurate	4 - Neither inaccurate nor accurate	5 - Slightly accurate	6 - Mostly accurate	7 - Completely accurate
a close up of a cake on a plate							
this is a photograph of a book on a table .							
this is a photograph of the <UNK> of the <UNK> .							
this is a photograph of a man in a suit and tie .							
this is a <UNK> photograph of an individual that by near .							
a close up of a pair of scissors on a table							

78. **Comments**

If you have any comments about this question, you can leave them here.

Image 39

79. *

Please indicate how accurate you think the following captions describe the image on a scale from 1 - completely inaccurate to 7 - completely accurate.
Mark only one oval per row.

	1 - Completely inaccurate	2 - Mostly inaccurate	3 - Slightly inaccurate	4 - Neither inaccurate nor accurate	5 - Slightly accurate	6 - Mostly accurate	7 - Completely accurate
a group of people sitting on top of a mountain .							
this is a view of a mountain range .							
this is a photograph of the <UNK> of the <UNK> , to .							
this is a <UNK> photograph of a man and a woman standing in front of a mountain in							
this is a <UNK> photograph of a single story frame building with a single story frame building with							
a view of a mountain range with mountains in the background .							

80. **Comments**

If you have any comments about this question, you can leave them here.

Image 40

81. *

Please indicate how accurate you think the following captions describe the image on a scale from 1 - completely inaccurate to 7 - completely accurate.
Mark only one oval per row.

	1 - Completely inaccurate	2 - Mostly inaccurate	3 - Slightly inaccurate	4 - Neither inaccurate nor accurate	5 - Slightly accurate	6 - Mostly accurate	7 - Completely accurate
a group of people standing in front of a building .							
this is a slide of a slide of a building with a sign on it .							
this is a slide of a slide of a tall brick building with a sign on it .							
this is a pole slide of the front facade of a toilet building . that by parking ,							
this is a pole slide of the front facade of a toilet building .							
a bird perched on top of a power line .							

82. **Comments**

If you have any comments about this question, you can leave them here.

Image 41

83. *

Please indicate how accurate you think the following captions describe the image on a scale from 1 - completely inaccurate to 7 - completely accurate.
Mark only one oval per row.

	1 - Completely inaccurate	2 - Mostly inaccurate	3 - Slightly inaccurate	4 - Neither inaccurate nor accurate	5 - Slightly accurate	6 - Mostly accurate	7 - Completely accurate
a group of men standing next to each other .							
men and women pose for a picture at a formal event .							
portrait view of the <UNK> , to .							
this is a <UNK> photograph of a man in a suit and tie .							
men and women pose in front of a group of men in suits and ties .							
a group of men standing next to each other .							

84.

Comments

If you have any comments about this question, you can leave them here.

Image 42

85. *

Please indicate how accurate you think the following captions describe the image on a scale from 1 - completely inaccurate to 7 - completely accurate.
Mark only one oval per row.

	1 - Completely inaccurate	2 - Mostly inaccurate	3 - Slightly inaccurate	4 - Neither inaccurate nor accurate	5 - Slightly accurate	6 - Mostly accurate	7 - Completely accurate
a statue of a bear sitting on a table .							
this is a slide of the front end of a toilet .							
this is a <UNK> slide of the front end of a toilet .							
this is a <UNK> slide of the front facade of a toilet building .							
this is a slide of the front facade of a toilet building with a covered porch . that							
a close up of a statue of a bear							

86. **Comments**

If you have any comments about this question, you can leave them here.

Image 43

87. *

Please indicate how accurate you think the following captions describe the image on a scale from 1 - completely inaccurate to 7 - completely accurate.
Mark only one oval per row.

	1 - Completely inaccurate	2 - Mostly inaccurate	3 - Slightly inaccurate	4 - Neither inaccurate nor accurate	5 - Slightly accurate	6 - Mostly accurate	7 - Completely accurate
a building with a sign on the front of it .							
this is a slide of a building with a sign on the side of the building .							
this is a pole slide of a tall building with a sign on it .							
this is a <UNK> photograph of a man and a woman pose in front of a toilet building							
this is a <UNK> photograph of a this is a pole slide of the front facade of a toilet brick building with a sign onman and a woman pose in front of a toilet building							
a black and white photo of a city street							

88.

Comments

If you have any comments about this question, you can leave them here.

Image 44

89. *

Please indicate how accurate you think the following captions describe the image on a scale from 1 - completely inaccurate to 7 - completely accurate.
Mark only one oval per row.

	1 - Completely inaccurate	2 - Mostly inaccurate	3 - Slightly inaccurate	4 - Neither inaccurate nor accurate	5 - Slightly accurate	6 - Mostly accurate	7 - Completely accurate
an old photo of a city street with many cars .							
this is a view of a bridge in the distance .							
this is a view of the city of the building .							
this is a <UNK> photograph of the building of the building . that by near , l .							
this is a <UNK> photograph of a large building in the background . that by near , l							
a black and white photo of a city street .							

90. **Comments**

If you have any comments about this question, you can leave them here.

Image 45

91. *

Please indicate how accurate you think the following captions describe the image on a scale from 1 - completely inaccurate to 7 - completely accurate.
Mark only one oval per row.

	1 - Completely inaccurate	2 - Mostly inaccurate	3 - Slightly inaccurate	4 - Neither inaccurate nor accurate	5 - Slightly accurate	6 - Mostly accurate	7 - Completely accurate
a group of people standing around a table .							
this is a photograph of a woman in a hospital room .							
a group of men and women standing in front of a building .							
interior view of the girl at the hospital .							
this is a <UNK> photograph of a man and a woman pose in front of a table .							
a woman standing in a kitchen next to a counter .							

92.

Comments

If you have any comments about this question, you can leave them here.

Image 46

93. *

Please indicate how accurate you think the following captions describe the image on a scale from 1 - completely inaccurate to 7 - completely accurate.
Mark only one oval per row.

	1 - Completely inaccurate	2 - Mostly inaccurate	3 - Slightly inaccurate	4 - Neither inaccurate nor accurate	5 - Slightly accurate	6 - Mostly accurate	7 - Completely accurate
a birthday cake that is on a table .							
this is a slide of a slide of a laptop computer .							
this is a slide of the front entrance of the building .							
this is a <UNK> photograph of a man and a woman in front of a laptop computer .							
this is a <UNK> slide of the front of a laptop . that by near , I .							
a close up of a keyboard and a mouse							

94. **Comments**

If you have any comments about this question, you can leave them here.

Image 47

95. *

Please indicate how accurate you think the following captions describe the image on a scale from 1 - completely inaccurate to 7 - completely accurate.
Mark only one oval per row.

	1 - Completely inaccurate	2 - Mostly inaccurate	3 - Slightly inaccurate	4 - Neither inaccurate nor accurate	5 - Slightly accurate	6 - Mostly accurate	7 - Completely accurate
a sign on a pole on a city street .							
this is a photograph of the front of the front of the building .							
this is a photograph of a man and a woman pose in front of a tall building with							
this is a <UNK> photograph of the front facade of the building . that by near , I							
this is a <UNK> photograph of the front facade of a toilet building with a sign on the							
a black and white photo of an old building							

96.

Comments

If you have any comments about this question, you can leave them here.

Image 48

97. *

Please indicate how accurate you think the following captions describe the image on a scale from 1 - completely inaccurate to 7 - completely accurate.
Mark only one oval per row.

	1 - Completely inaccurate	2 - Mostly inaccurate	3 - Slightly inaccurate	4 - Neither inaccurate nor accurate	5 - Slightly accurate	6 - Mostly accurate	7 - Completely accurate
a man and a woman are standing under an umbrella .							
this is a portrait of a man and a woman pose in front of a building .							
this is a portrait of a man and a woman pose in front of a porch .							
this is a <UNK> glass plate of a man and a woman pose in front of a toilet							
this is a <UNK> photograph of a man and a woman pose in front of a toilet building							
a black and white photo of a man and a woman							

98.

Comments

If you have any comments about this question, you can leave them here.

Image 49

99. *

Please indicate how accurate you think the following captions describe the image on a scale from 1 - completely inaccurate to 7 - completely accurate.
Mark only one oval per row.

	1 - Completely inaccurate	2 - Mostly inaccurate	3 - Slightly inaccurate	4 - Neither inaccurate nor accurate	5 - Slightly accurate	6 - Mostly accurate	7 - Completely accurate
a group of people sitting at a table .							
a woman sitting on a chair with a teddy bear .							
this is a portrait of a man and a woman in front of a table .							
a man and woman pose in front of a table in a room .							
a man and woman pose in front of a table .							
a black and white photo of a woman and a child							

100. **Comments**

If you have any comments about this question, you can leave them here.

Image 50

101. *

Please indicate how accurate you think the following captions describe the image on a scale from 1 - completely inaccurate to 7 - completely accurate.
Mark only one oval per row.

	1 - Completely inaccurate	2 - Mostly inaccurate	3 - Slightly inaccurate	4 - Neither inaccurate nor accurate	5 - Slightly accurate	6 - Mostly accurate	7 - Completely accurate
a black and white photo of a city street .							
a group of men and women standing in front of an airplane .							
a group of men and women standing in front of a building .							
view of the <UNK> of the girl , to .							
a group of men and women stand in front of an airplane .							
a black and white photo of a group of people on a boat .							

102.

Comments

If you have any comments about this question, you can leave them here.

Image 51

103. *

Please indicate how accurate you think the following captions describe the image on a scale from 1 - completely inaccurate to 7 - completely accurate.
Mark only one oval per row.

	1 - Completely inaccurate	2 - Mostly inaccurate	3 - Slightly inaccurate	4 - Neither inaccurate nor accurate	5 - Slightly accurate	6 - Mostly accurate	7 - Completely accurate
a group of stuffed animals sitting on top of a table .							
this is a slide of the front door of the front door .							
this is a portrait of a man in a suit and tie .							
this is a <UNK> photograph of an individual in a room .							
this is a row slide of the front facade of a toilet building .							
a painting of a man in a suit and tie .							

104. **Comments**

If you have any comments about this question, you can leave them here.

105. **Please indicate what you think accuracy of the images referred to in this questionnaire?** *

An image caption's accuracy refers to the extent with which the caption
Check all that apply.

- [ ] describes what is depicted in the image
- [ ] is grammatically correct
- [ ] contains more than five words
- [ ] all of the above