Generating Natural Language Summaries for Image Sets

by

Akash Abdu Jyothi

Dual Degree (B. Tech. and M.Tech.), Indian Institute of Technology Madras, 2013

Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of Master of Science

in the School of Computing Science Faculty of Applied Sciences

©Akash Abdu Jyothi 2018
SIMON FRASER UNIVERSITY Summer 2018

Copyright in this work rests with the author. Please ensure that any reproduction or re-use is done in accordance with the relevant national copyright legislation.
Approval

Name: Akash Abdu Jyothi

Degree: Master of Science (Computing Science)

Title: Generating Natural Language Summaries for Image Sets

Examining Committee: Chair: Parmit Chilana
Assistant Professor

Greg Mori
Senior Supervisor
Professor

Anoop Sarkar
Supervisor
Professor

Maxwell Libbrecht
Internal Examiner
Assistant Professor

Date Defended: May 31, 2018
Abstract

We address the problem of summarizing an image set with a natural language caption. We present PlacesCap, a new dataset for image set summarization. Our dataset consists of 11,661 image sets with a total of 116,113 images, where each set is summarized by a 3 sentence caption. We propose novel pooling operators for permutation invariant sets of feature maps, and empirically evaluate image set summarization models based on those operators. We also conduct experiments of image set classification and show competitive performance for the proposed set pooling operators.

Keywords: image set summarization; natural language summary generation; set compression; set pooling
Acknowledgements

First and foremost, I would like to thank my supervisor Dr. Greg Mori for envisioning this project, and providing constant support and encouragement throughout its course. His guidance was critical for completion of this work and immensely instructive for my learning. I am deeply grateful to Fred for his involvement in this project from the very beginning. He was the go-to person whenever I was stuck with a problem, and his advice always cleared things up for me. I would like to thank Thibaut for his active participation in the project with many stimulating discussions, and his experiments that helped us understand the dataset better. I would also like to thank Dr. Anoop Sarkar and Dr. Leonid Sigal for the many insightful discussions over this project.

I would like to thank to Datapure for their prompt data annotation services which helped us create our dataset without major delays.

I am grateful to my friends Nagender, Sreenath, Josh, Srikanth, Nishant, Pratik, Aniket, Sha, Nazanin and many others for keeping me grounded during the difficult process of growing up. To the authors of all the books that I read in the last two years, I can’t name all of you here but thanks a ton. For showing me the possibilities of being through wonderful music that I barely understand, I am forever indebted to Mr. Johann Sebastian Bach.

I would like to express my gratitude to my parents, who always supported me in my endeavours whether they saw any sense in it or not. Finally, I would like to thank my sister Sangeetha for her proactive support and invaluable guidance in life, the universe and everything.
# Table of Contents

Approval ................................................................. ii

Abstract ................................................................. iii

Acknowledgements ....................................................... iv

Table of Contents ....................................................... v

List of Tables ............................................................ vii

List of Figures ........................................................... viii

1 Introduction ................................................................ 1
   1.1 Contributions ...................................................... 2

2 Related Work ............................................................ 4
   2.1 Image set summarization as selecting exemplars .......... 4
   2.2 Summarization of photo streams .............................. 5
   2.3 Visual captioning ................................................... 5
   2.4 Visual Paragraph Generation .................................... 6

3 Dataset ....................................................................... 8
   3.1 Image Set Collection ............................................... 8
   3.2 Image Set Annotation .............................................. 9

4 Models ....................................................................... 11
   4.1 Set Summarization Model ........................................ 11
   4.2 Pooling Techniques ................................................ 11
      4.2.1 Set Aware Pooling ............................................ 13
      4.2.2 Set Mask Pooling ............................................. 14
      4.2.3 Set Weight Pooling .......................................... 15
      4.2.4 Generalized Mean Pooling ................................. 15

5 Experiments and Results ............................................... 16
### List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table 3.1</td>
<td>Distribution of categories in PlacesCap dataset</td>
<td>9</td>
</tr>
<tr>
<td>Table 5.1</td>
<td>Analysis of predicted captions on the validation set that has 2332 image sets. Learning rate was set to 0.001. The values are reported for models from training epoch 18 (chosen based on peak performance of most of the models).</td>
<td>18</td>
</tr>
<tr>
<td>Table 5.2</td>
<td>Classification accuracy obtained using mean pooling and the proposed pooling methods.</td>
<td>18</td>
</tr>
<tr>
<td>Table 5.3</td>
<td>Summary of the results obtained using mean pooling and the proposed pooling methods. B-n = BLEU-n, Perp. = Perplexity, CE = Cross Entropy.</td>
<td>19</td>
</tr>
</tbody>
</table>
# List of Figures

| Figure 1.1 | A set of images showing a tourist location with a summary description below. | 2 |
| Figure 2.1 | Summary of a set of 2000 images of Vatican. Image from Simon *et al.* [22]. | 5 |
| Figure 2.2 | Figure from Kim *et al.* [12] showing (a) photo stream input from multiple users and (c) output storyline graph with time stamps. (b) Optional input of friendship graph is utilized for creating weakly-personalized storylines by giving higher weightage for photo streams from close friends. | 6 |
| Figure 2.3 | Figure from Xu *et al.* [29] depicting image captioning model with visual attention. | 6 |
| Figure 2.4 | Figure from Krause *et al.* [14] depicting a hierarchical recurrent neural network used for describing an image with a paragraph. | 7 |
| Figure 3.1 | Sample image sets, one from each category in the dataset, along with annotated summary caption. | 10 |
| Figure 4.1 | Schematic of the proposed set summarization model. | 12 |
| Figure 4.2 | Schematic of the proposed set pooling operations. | 14 |
| Figure 5.1 | Results from the experiment where we use top 1, 2, 3, 4, 5 and all images respectively from the image sets for a summarization model that uses mean pooling of image features. | 20 |
| Figure 5.2 | Language metrics and cross entropy on validation set for a mean pooling based model across training epochs for beam sizes 1(sampling), 3, 5 and 7. Higher values indicate better performance for CIDEr and BLEU-4. | 21 |
| Figure 5.3 | Summary prediction examples from training set. GT = ground truth summary, Pred = predicted summary. | 22 |
| Figure 5.4 | Summary prediction examples from validation set. GT = ground truth summary, Pred = predicted summary. | 23 |
Chapter 1

Introduction

Comprehending data collections on the internet is often difficult because of their sheer volume. These collections may consist of images (Flickr sets), audio (Spotify playlists), video (YouTube playlists) or any combination of these along with text (TripAdvisor, Airbnb, Amazon listings). Categorization and tagging of content is helpful but they fall short of capturing details that will enable fine search. Fig. 1.1 shows a motivational example of an image set of a tourist location, and its summary caption that captures details that may be of interest to a potential tourist.

Many professional and scientific disciplines also currently require humans to manually sift through large collections of images to obtain the required information. For example, a search and rescue effort may involve volunteers combing through satellite imagery for signs of a missing person, a humanitarian disaster relief project may require volunteers to assess the extent of damage across a large area so that resources can be appropriately allocated, or a medical diagnosis may require a practitioner to gather information from across many image scans.

We focus on the problem of producing meaningful textual output - summaries in natural language - to summarize a large set of images. Compact set summaries will assist users to navigate easily to interesting content in large data collections, and enable users to tell at a glance whether particular sets are relevant to their goals.

Automatic understanding of images is a common objective in computer vision. Most image understanding tasks, such as object detection, semantic segmentation, image captioning, and visual question answering, take a single image as input and infer task-specific information from the image. For example, segmenting an image into its semantic categories, or answering a question about a single image. In contrast, image set summarization involves understanding an unstructured collection of images. We focus on summarization of image sets, though our model can be easily extended to summarize other input modalities by appropriate replacement of the image feature extraction module.

Previous works on image set summarization [30, 24, 22, 23, 27] focus on selecting summary images from the set. This style of summarization is useful in certain cases like sum-
This is a sandy beach. The beach is crowded with many tourists. There are rock formations and grassy patches of land nearby.

Figure 1.1: A set of images showing a tourist location with a summary description below.
• We provide a new dataset PlacesCap for image set summarization. PlacesCap contains 11,661 image sets of popular tourist locations around the world, with 116,113 images in total. Every set is summarized by a 3 sentence description.

• We propose novel parametrised set pooling operators, that are generic and learned through backpropagation. We show effectiveness of the proposed operators through experiments on image set classification.
Chapter 2

Related Work

There has been a great interest in learning to describe visual information using natural language in the recent years. To our knowledge, our work is the first attempt to describe permutation invariant image sets with a multi sentence caption. In the following sections, we review prior work from research areas that are closely related to this task. While an extensive review of works in this domain is beyond the scope of this document, we categorize them into a few major directions of research and discuss major works in those areas.

2.1 Image set summarization as selecting exemplars

Over the years, there have been many attempts to summarize image sets by selecting representative images from the set. The seminal work by Simon et al. [22] attempt scene summarization task on multi-user collection of online images (eg: a collection of images pertaining to a tourist location). Their method operates over sets that have thousands of images by first clustering the set into groups of related images based on SIFT [18] feature co-occurrences. Canonical views are then selected from the groups based on a likelihood measure which again uses feature co-occurrences, but within the group. They also select a representative tag for each canonical view based on user tags associated with each image. An example of summary images is given in Fig. 2.1.

Yang et al. [30] treat summarization as a dictionary learning problem where the objective is to learn a small subset of the image collection from which one can reconstruct most of the images in the set. Xu et al. [27] propose a method for selecting exemplars which considers both images and associated tags. They extend the affinity propagation [7] algorithm so as to predict summary images using heterogeneous data consisting of images and tags.

Tschiatschek et al. [24] optimize a learning objective over a mixture of submodular functions to arrive at a subset of images that summarizes a larger set. They use human annotated summary image sets for constructing a loss function. This line of work was later extended by Singla et al. [23] to learn exemplars through noisy evaluation of summary sets,
like click or rating based metrics from user feedback guiding the optimization instead of human annotated summary image sets.

2.2 Summarization of photo streams

Photo streams form a subset of online image collections, and they are distinct from image sets we consider in this work in that they have temporal ordering. Kim et al. [12] propose a method for creating a storyline graph given a set of photo streams related to a particular topic. An example of a summary storyline is shown in Fig. 2.2.

Another work in this domain is by Kim et al. [11] where photo streams and blog posts with images are jointly aligned for summarizing tasks. Blog posts were found to be beneficial in the tasks of automatically assigning titles and location information for images in photo streams. Photo streams, on the other hand, were useful in interpolating between the images in blog posts.

2.3 Visual captioning

Image captioning aims to describe the contents of an image with a natural language sentence. This task has become popular in the recent years particularly with the introduction of large image captioning datasets like Flickr30k [31] and MSCOCO [17] where images are annotated by crowd workers. Encoder-decoder style models for image captioning, first proposed by Vinyals et al. [26], is currently the most successful framework for solving this task. This framework uses a recurrent neural network language model which takes feature representation of the image as input to predict words one by one.
Figure 2.2: Figure from Kim et al. [12] showing (a) photo stream input from multiple users and (c) output storyline graph with time stamps. (b) Optional input of friendship graph is utilized for for creating weakly-personalized storylines by giving higher weightage for photo streams from close friends.

Figure 2.3: Figure from Xu et al. [29] depicting image captioning model with visual attention.

Xu et al. [29] propose a major modification over the simple captioning model by adding visual attention mechanism for predicting the caption. As depicted in Fig. 2.3, this allowed the model to focus on different areas of the image while generating different words. Lu et al. [19] extend spatial attention model by adding a temporal attention framework that allows the language model to choose when to attend to the image.

MSR-Video To Text [28] is a large scale dataset for video captioning. It has 10k web clips where each clip is described by about 20 different sentences. They provide baseline results using a similar modeling techniques as Vinyals et al. [26], focusing primarily of video feature extraction methods.

2.4 Visual Paragraph Generation

Next we discuss recent works that attempt to generate a paragraph description for visual information — images, sequence of images and videos. Krause et al. [14] present a dataset
for image paragraph generation by annotating a subset of MS COCO [17] images with paragraphs. Their proposed model, as shown in Fig. 2.4 comprises of a region of interest detector, whose features are pooled and passed to a language model. The language model is a hierarchical recurrent neural network with two stages, first stage predicting sentence codes and the second stage decoding it word by word.

Liang et al. [16] propose a method for generating paragraph descriptions for images in a semi-supervised manner by using image-sentence pairs and standalone text paragraphs that are not paired with images. They use the same region of interest detector as in Krause et al. [14], and a hierarchical language model that takes cues from an adversarial discriminator that assesses plausibility of sentences and change of topic across sentences. The adversarial discriminator uses standalone text paragraphs to distinguish between real and synthesized paragraphs.

Park et al. [21] propose the task of retrieving a sequence of sentences from a dataset of blog posts to describe a photo stream. Though this work has many similarities to ours, there is significant difference in that temporal ordering of input photo stream dictates content and order of sentences in their output. In our task, we explicitly focus on creating summary captions that does not depend on the order of images in the input set. While both these tasks may be applicable in many scenarios, depending on the context of application, one of these tasks may have more relevance than the other. Huang et al. [10] target a similar problem and introduce a dataset with sequence of images each paired with a sentence caption. They provide three versions of captions for each image sequence that vary in the level of smoothness in transition across sentences — from disconnected sentences to a story.

Yu et al. [32] target the problem of generating paragraph descriptions for videos. They use a hierarchical RNN model with attention for this purpose.

Figure 2.4: Figure from Krause et al. [14] depicting a hierarchical recurrent neural network used for describing an image with a paragraph.
Chapter 3

Dataset

In this section, we describe details about PlacesCap dataset. It consists of 11,661 image sets across 8 categories, with a total of 116,113 images. Each set contains up to 10 images of a popular tourist attraction. Table 3.1 lists all the categories and distribution of images across them. Every image set has a three-sentence summary caption describing the place as seen in the images. Summary caption captures nature of the location (often same as its category name), and also describes major attractions in the location. We believe a summary caption of this style would benefit tourists who are exploring potential holiday options. Some samples from the dataset are shown in Fig. 3.1.

After defining the task, we created the dataset in two stages - image set collection, and image set annotations.

3.1 Image Set Collection

We used Google image search for obtaining image sets of tourist locations. We constructed search phrase by using place name followed by name of the city where it is present. For example, Big Ben in London gives us the search phrase "big ben london".

To create a list of place names, we first collected names of more than 400k tourist locations around the world from popular travel websites. We then grouped the tourist attractions into different categories based on tags that were associated with them. Out of ~200 tags, we manually chose 8 tags as categories for our dataset. Categories were chosen based on quality of image sets that it contained with respect to this task — coherence of top Google image results with respect to the place name, and the ease with which a human could provide a meaningful summary caption that corresponds to information in the image set. We selected 12,500 locations from across 8 categories and downloaded top 10 results from Google image search for each of those locations.
<table>
<thead>
<tr>
<th>Category</th>
<th>Number of Image Sets</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parks</td>
<td>2,548</td>
</tr>
<tr>
<td>Nature &amp; Wildlife Areas</td>
<td>1,912</td>
</tr>
<tr>
<td>Bodies of Water</td>
<td>1,899</td>
</tr>
<tr>
<td>Gardens</td>
<td>1,454</td>
</tr>
<tr>
<td>Hiking Trails</td>
<td>1,446</td>
</tr>
<tr>
<td>Mountains</td>
<td>928</td>
</tr>
<tr>
<td>Beaches</td>
<td>739</td>
</tr>
<tr>
<td>Waterfalls</td>
<td>735</td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td><strong>11,661</strong></td>
</tr>
</tbody>
</table>

Table 3.1: Distribution of categories in PlacesCap dataset

### 3.2 Image Set Annotation

We outsourced the task of writing summary caption for each image set to DataPure [2], after a failed attempt to crowdsourcethe task through Hive [4]. The primary difficulty in crowdsourcing the task through Hive, or other crowdsourcing platforms like Figure Eight [3] or Amazon Mechanical Turk [1], was the absence of a platform which can handle sets of images at a time. Platforms that focus on getting annotations from crowd workers typically have templates that support popular annotation tasks that involve single images. Unlike other data annotation platforms, DataPure operates with a group of workers who are also full time employees in the company. This enabled us to get image sets summarized by a controlled group of workers. We used Google Drive as the platform to share and view image sets.

During annotation of an image set, a worker was shown all images from the set along with the name of the location and asked to write a 3 sentence caption. The worker was directed to first write about the location, followed by description of major attractions in that location, all of which could be inferred given the images. Workers were asked to skip the image sets that do not have coherent information about the location. After annotation, we were left with 11,661 sets in the final dataset.

Once all the sets were annotated, a second round of quality check was done to reduce error in captions. In this stage workers were shown the image set along with the caption, and asked to modify the caption for grammatical and other errors.

A notable challenge arose in the annotation stage during worker training phase. The particular group of workers assigned to this task was accustomed to a style of writing which involved extravagant use of adjectives and imagery. Because of this habit, even after providing clear written guidelines (Appendix A), initial round of summary captions contained literary embellishments that were unacceptable from the point of view of our task. We conducted multiple online training sessions directly with the workers to teach them the objective of the annotation task.
Beaches: This is a calm beach. There is a lighthouse on the beach. The beach has a wooden deck for walking.

Gardens: This is a delightful garden. Pink flowers surround the garden. An arched walkway is a present in the garden.

Hiking Trails: This is a hard hiking trail. The forest trees surround the hiking trail. There is a huge statue by the trail.

Mountains: This is a beautiful mountain. A waterfall tumbles down the mountain. There is a sculpture carved out of the rocks.

Nature & Wildlife Areas: This is a campsite. There are cottages amid greenery on the hill by the sea. There are stunning walkways in the park.

Parks: This is a park for recreational activities. A merry go round swing is present in the park for the kids. There is a circular shelter in the park with the conical roof.

Bodies of Water: This is a beautiful lake surrounded by mountain ranges. Rocks surround the area. There is a wooden deck near the lake.

Waterfall: This is a mighty waterfall on the rocky mountain. There is an opportunity to swim in the waterfall. There is a riverside space to watch waterfall falling.

Figure 3.1: Sample image sets, one from each category in the dataset, along with annotated summary caption.
Chapter 4

Models

Vinyals et al. [26] first proposed the encoder-decoder based model for image captioning task. In the original framework, input image is encoded into a feature representation using a convolutional neural network, which is then decoded into a caption one word at a time by a recurrent neural network. The model is trained end-to-end by using word level cross entropy between the predicted sentence and the ground truth sentence as the loss function. This model structure has become quite popular with many captioning projects [29, 19, 14, 16, 21, 32, 28] adopting this framework. In this section, we discuss our set summarization model, which is based on the same idea, with emphasis on pooling operation for sets.

4.1 Set Summarization Model

We follow an encoder-decoder approach similar to Vinyals et al. [26] for our model. Fig. 4.1 depicts the proposed image set summarization model. Given a set of N images \( I = \{I_1, I_2, \ldots, I_N\} \), the model predicts summary caption \( y \).

The encoder consists of Resnet-152 [8] followed by a fully connected layer \( fc_{img} \). We use 2048-dimension features before the final fully connected layer from Resnet-152 pretrained on Imagenet [6]. Resnet-152 features of each image is passed through \( fc_{img} \) to get the set of image features \( X = \{x_1, x_2, \ldots, x_N\} \). We use pooling operation on this set and pass the pooled features through another fully connected layer \( fc_{set} \).

Output of \( fc_{set} \) is fed into an LSTM [9] language model to predict \( y \). The language model predicts word tokens for all the three sentences in sequence, including the tokens representing period character at the end of each sentence.

4.2 Pooling Techniques

We experiment with different pooling operations in our summary caption generation model. A natural choice is global average (or max) of all individual image features to get a single feature for the image set. We observe that mean of a set of feature maps considers each
This is a sandy beach. There are a few travellers on the beach. The region is surrounded by rocky hills. 

Figure 4.1: Schematic of the proposed set summarization model
feature in a manner that is agnostic to the global information in the entire set. We believe this is a significant limitation of simple pooling operations like mean or max.

Depending on the context of task, information that needs to be pooled from each member of a set may depend on the entire contents of set. For set summarization task, suppose we have a sample set which has an image of a beach with some beach chairs and a covered rest area. The model may be required to focus on any one of those three items or some combination of the three depending on information present in the set of images.

We hypothesize that feature maps obtained by appending information about the full set to each image feature, followed by a fully connected layer, achieves the above objective. Following this idea, we propose three novel parameterized pooling operators that can selectively choose information from the individual feature maps based on global set information. The proposed operators are permutation invariant, and computation scales linearly with number of features $N$ in the set. Parameters of these pooling operators are learned through backpropagation. Thus, they are generic operators for similar models that require aggregation of sets on features.

Apart from the above ideas, we also conduct experiments by using generalized mean pooling. We did not conduct extensive experiments using max based pooling operators since our experiments showed higher performance using mean pooling. We describe the three proposed models, given a set $\mathbf{X} = \{x_1, x_2, \ldots, x_N\}$, we first learn a new set $\hat{\mathbf{X}} = \{\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_N\}$. Fig. 4.2 depicts the learning of new set. Once we have $\hat{\mathbf{X}}$, we compute global average of its members to obtain pooled feature.

### 4.2.1 Set Aware Pooling

Given a set $\mathbf{X}$ of features $x_i$, we concatenate each feature of the set with $\text{mean}(\mathbf{X})$. Each concatenated feature is passed through fully connected layer $f_{c_a}$ to obtain a new set $\hat{\mathbf{X}}$ of features $\hat{x}_i$. We compute global average of members of $\hat{\mathbf{X}}$ to obtain a pooled feature $x_a$. We call this operation set aware pooling.

\[
\hat{x}_i = W_a \left( x_i \sim \left( \frac{1}{N} \sum_{j=1}^{N} x_j \right) \right) + b_a
\]

\[
x_a = \frac{1}{N} \sum_{i=1}^{N} \hat{x}_i
\]

(4.1)

where $W_a$ and $b_a$ are the parameters of $f_{c_a}$. The operator $\sim$ represents concatenation.
4.2.2 Set Mask Pooling

Given a set $\mathcal{X}$ of features $x_i$, we concatenate each feature of the set with $\text{mean}(\mathcal{X})$. Each concatenated feature is passed through fully connected layer $f_{c_m}$ followed by a sigmoid function to obtain a set of masks $m_i$. We perform Hadamard product of mask $m_i$ with corresponding $x_i$ to get a new set $\hat{\mathcal{X}}$ of features $\hat{x}_i$. We compute global average of members of $\hat{\mathcal{X}}$ to obtain a pooled feature $x_m$. We call this operation set mask pooling.

$$m_i = \text{sigmoid} \left( W_m \left( x_i \sim \left( \frac{1}{N} \sum_{j=1}^{N} x_j \right) \right) + b_m \right)$$

$$\hat{x}_i = m_i \cdot x_i$$

$$x_m = \frac{1}{N} \sum_{i=1}^{N} \hat{x}_i$$

where $W_m$ and $b_m$ are the parameters of $f_{c_m}$. The operator $\sim$ represents concatenation and the operator $\cdot$ represents Hadamard product.
4.2.3 Set Weight Pooling

This is similar to set mask pooling, with the only difference that we learn a scalar \( m_i \), instead of a vector, which is used to weigh each image feature while computing the pooled feature \( x_w \).

\[
m_i = \text{sigmoid} \left( W_w \left( x_i \sim \left( \frac{1}{N} \sum_{j=1}^{N} x_j \right) + b_w \right) \right)
\]

\[
\hat{x}_i = m_i x_i
\] (4.3)

\[
x_w = \frac{1}{N} \sum_{i=1}^{N} \hat{x}_i
\]

where \( W_w \) and \( b_w \) are the parameters of a fully connected layer \( fc_w \). The operator \( \sim \) represents concatenation.

4.2.4 Generalized Mean Pooling

Generalized mean or power mean is a set aggregating function whose special cases include geometric, harmonic and arithmetic means. Generalized mean \( x_g \) of a set \( \{x_1, x_2, \ldots, x_N\} \) is given by,

\[
x_{g_k} = \left( \frac{1}{p_k} \sum_{i=1}^{N} x_{ik}^{p_k} \right)^{\frac{1}{p_k}}
\] (4.4)

where \( p \) is a vector of the same size as \( x_i \), allowing us to learn separate parameters for individual elements in the features vector. Index \( k \) denotes elements in the vector, and takes values from 1 to size of the feature vector \( x_i \).
Chapter 5

Experiments and Results

We conduct experiments on image set summary generation using PlacesCap. To demonstrate the necessity of using image sets for this task, we conduct experiments where we progressively restrict the number of images from the set available to a simple mean pooling based model. We also conduct experiments using the proposed set pooling methods on a simpler classification task designed using the same dataset, and we show competitive performance of set mask pooling over simple mean pooling operation.

5.1 Experimental Setup

In this section, we discuss details about data preprocessing, implementation of models, and evaluation metrics that are used in the experiments.

5.1.1 Data Preprocessing

We divide the dataset consisting of 11,661 image sets into train, validation and test sets in 60:20:20 ratio. For each summary caption that consists of three sentences, we add a \texttt{<START>} token to the beginning and a \texttt{<STOP>} token at the end after the period character. Vocabulary size for this dataset is 3971.

5.1.2 Implementation details

\textbf{Encoder.} Given a set of images, each image is first resized to 224*224 and input to a Resnet-152 model pretrained on Imagenet. We do not finetune Resnet-152 model. Resnet features of dimension 2,048 is passed through fully connected layer $f_{c_{img}}$ giving features of dimension 1,024. These features are pooled using one of the set pooling methods and then passed through fully connected layer $f_{c_{set}}$ to get image set features that are of the word embedding size, 256. We apply batch normalization over this layer with momentum (weight on the moving average estimate used for normalization) set to 0.01. All weights in fully connected layers in the encoder are initialized as normal distribution having 0 mean and 0.02 standard deviation, with bias terms initialized to 0.
**Decoder.** Set pooled feature from the encoder is passed to an LSTM language model. We use 256 as word embedding size and 512 as LSTM hidden size. Dropout with a probability 0.2 is applied on input to the LSTM. LSTM output is passed through a fully connected layer $f_{vocab}$ which gives a feature of length vocabulary size. Weights of $f_{vocab}$ and embedding dictionary layer is initialized with a uniform distribution in the range $[-0.1, 0.1]$, with bias terms initialized to 0.

We use word level cross entropy as loss function and optimize the model with Adam [13]. Learning rate is set to 0.0005, with a decay of 0.5 at epochs 8, 16, 24, 32, 40 and 48. We initially chose teacher forcing methodology for training, where input to the language model at every time step is the ground truth word token from the previous step. Since this training strategy was not found to be effective, we opted for a scheduled sampling [5] approach where we use output word from the language model as input to the next step with some probability. We used inverse sigmoid decay for sampling probability $p_i$ across epochs as given in the below equation.

$$p_i = (1 - \epsilon_i)^2$$

$$\epsilon_i = \frac{k}{k + \exp(i/k)}$$

(5.1)

where $i$ is the epoch and $k = 90$. While scheduled sampling gave more varied captions, it also led to two observable negative effects in terms of sentence quality. Firstly, it resulted in instances of words or the period character being repeated. Secondly, it resulted in incomplete sentences. The following caption captures both these defects.

This is a waterfall waterfall. The forest. The swimming in the river stream invites the crowd. There is a hiking trail by the waterfall.

Table 5.1 shows the effect of maximum scheduled sampling probability on the number of repeated tokens. Based on this experiment, we chose to restrict maximum sampling probability to 0.25.

We select best performing model from the training epoch where combined CIDEr and METEOR score on the validation set is highest. During inference, we perform sampling using the decoder to predict summary caption.

**5.1.3 Evaluation Framework**

We evaluate predicted summary captions based on CIDEr [25], BLEU [20] and METEOR [15] scores. We use full summary caption with all three sentences including period characters for measuring prediction quality with these language metrics. This is important in this task
### Table 5.1: Analysis of predicted captions on the validation set that has 2332 image sets.

Learning rate was set to 0.001. The values are reported for models from training epoch 18 (chosen based on peak performance of most of the models).

<table>
<thead>
<tr>
<th>Max. Schedule Sampling Probability</th>
<th>No. of Unique Captions</th>
<th>No. of Consecutive Repeated Tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.20</td>
<td>991</td>
<td>181</td>
</tr>
<tr>
<td>0.25</td>
<td>1097</td>
<td>164</td>
</tr>
<tr>
<td>0.30</td>
<td>1238</td>
<td>297</td>
</tr>
<tr>
<td>0.50</td>
<td>1221</td>
<td>972</td>
</tr>
</tbody>
</table>

### Table 5.2: Classification accuracy obtained using mean pooling and the proposed pooling methods.

<table>
<thead>
<tr>
<th>Pool Type</th>
<th>Train %</th>
<th>Val %</th>
<th>Test %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>66.04</td>
<td>60.46</td>
<td>60.48</td>
</tr>
<tr>
<td>Set Mask</td>
<td>65.24</td>
<td>62.48</td>
<td><strong>62.02</strong></td>
</tr>
<tr>
<td>Set Aware</td>
<td>67.78</td>
<td>59.86</td>
<td>59.92</td>
</tr>
<tr>
<td>Set Weight</td>
<td>63.78</td>
<td>61.19</td>
<td>59.84</td>
</tr>
</tbody>
</table>

since language model is expected to learn the regularities not just within sentences, but also across different sentences.

### 5.2 Experiment on Classification

To evaluate the proposed pooling methods in a simpler setting, we conduct experiments on classification using the same dataset. For this purpose, we use category of each image set as its class label. Classification model is mostly similar to the encoder in summarization model with the main difference that $f_{set}$ now gives an output feature of dimension 8 (number of classes). Applying softmax function over this feature gives predicted probability of the classes. Another difference between this model and encoder in summarization model is that we use dropout with probability 0.2 on the set pooled features. Classification model is trained using SGD with a learning rate of 0.005. We use momentum of 0.9 for SGD.

Results on this experiment are shown in Table 5.2. We found that the model based on generalized mean pooling was difficult to train because of exploding gradients problem. We report results for all other pooling operations. We can see that set mask pooling achieves ~1.5% better test accuracy compared to mean pooling.

### 5.3 Dataset Analysis

Datasets in computer vision are associated with specific tasks so that research community can devise models that solve that task. Even when they are curated manually, vision datasets
Table 5.3: Summary of the results obtained using mean pooling and the proposed pooling methods. B-n = BLEU-n, Perp. = Perplexity, CE = Cross Entropy.

Based on real world data suffer from having strong priors, sometimes unrelated to the proposed task, that can be exploited by models to achieve high performance on the dataset.

We conduct experiments to validate the necessity of having full set of images in our dataset for predicting summary caption. To this end, we run a mean pooling based summary generation model using subsets of the entire image set as input. Our dataset has up to top-10 Google image search results in each set. A natural choice for subset of images is to select top-k images, with k<10, for each experiment. We conduct experiments with k=1, 2, 3, 4, 5 and also with all the images in the set. Fig. 5.1 shows BLEU-4 and CIDEr scores obtained for different inputs. We can see that there is a difference of ~2 BLEU-4 score and ~15 CIDEr score between the models that use only 1 image and the full set of images respectively.

5.4 Results

Table 5.3 shows the results of our experiments on summary generation. Even though we found notable difference between set mask pooling and mean pooling in classification task, all the pooling operations gave very similar results in summarization task. To explore this further, we performed inference using beam search with different beam sizes. Results from beam search experiment are shown in Fig. 5.2.

5.4.1 Qualitative Results

We present randomly selected sample image sets with predicted summary captions as qualitative results to give the reader a better understanding of model performance. Fig. 5.3 and Fig. 5.4 shows predictions and ground truth summary captions from training set and validation set respectively for mean pooling based model.

5.5 Analysis of Results

Our trained models mostly predict grammatically correct sentences that appear acceptable. Though many samples from qualitative results look reasonable from a human perspective, often they don’t match well with the ground truth. We do not obtain significant improvement...
Figure 5.1: Results from the experiment where we use top 1, 2, 3, 4, 5 and all images respectively from the image sets for a summarization model that uses mean pooling of image features.
in conventional sentence evaluation metrics — BLEU, CIDEr and METEOR. We note that these metrics are based on precision and/or recall of n-grams with respect to ground truth, and give equal weight to all the words in a sentence. We believe this may be a limitation in evaluating model performance for this task for two reasons.

- Summary captions in this dataset were designed to follow a prescribed format, and focus on the few major items in the image sets. Annotated captions contain many words ("the", "there", "this", "has", "a", "is", "are", "region" etc.) that are necessary for the caption, but not relevant to the information that is parsed from image set. They also have many adjectives that are subjective. Presence of these words could be one reason for similar performance of all the models in evaluation.

- Image sets often contain many scenes with a large number of objects. Based on subjective evaluation of summary prediction results, we observe that an image set could be described in a number of ways even when constrained by rules like those we imposed for annotating this dataset (Appendix A).

The above factors point towards the need for a more sophisticated evaluation metric for this task. If we choose to overlook these factors, the fact that beam search gives lower performance than sampling strongly indicates that the model requires more/better quality data to learn regularities across sentences, and give more accurate predictions.
Figure 5.3: Summary prediction examples from training set. GT = ground truth summary, Pred = predicted summary.
Figure 5.4: Summary prediction examples from validation set. GT = ground truth summary, Pred = predicted summary.
Chapter 6

Conclusion

We presented a supervised learning framework that learns to predict summary caption for a set of images. We proposed novel parameterised set pooling operators for this purpose, that also generalize beyond this task. We created PlacesCap, a dataset for image set summarization, and conducted experiments on summary generation and classification. Experiments on classification indicate that some of the proposed pooling operators perform better than mean pooling. Qualitative results on summarization demonstrate that the proposed model framework is able to predict image set summaries with some success. Quantitative results on summarization indicate similar performance for all the model variants. We conclude by discussing limitations and future work.

6.1 Limitations and Future Work

PlacesCap dataset and the proposed models is a first step towards solving the task of generating natural language summary for image sets. While we focused on a supervised learning framework, we believe there is great potential for a semi-supervised approach using freely available information from the web.

Summarization is a subjective task and the variety of ways in which an image set can be summarized is many fold. We tried to restrain the subjectivity by limiting the scope of summary. We believe this solution could be refined further with better annotation instructions and quality control methods. The problem can also be made more tractable to evaluate if dataset summary captions have relatively more words that are directly relevant to the image set.

Finally, evaluation metrics that are currently in use overlook the subjective nature of summaries. While this can be partially overcome by collecting more data per image set to get a diverse set of summaries, developing more sophisticated evaluation frameworks that go beyond using only the ground truth annotations would be an interesting line for future work.
Bibliography


[3] Figure eight. https://www.figure-eight.com/.


Appendix A

Labelling Guidelines for Annotators

We present below the document that was shared with Datapure for training the annotators. While the annotators were fluent in English, they were not native English speakers. This prompted us to provide a short dictionary of common scenes and objects from our dataset, that they may not be familiar with.

While this document was shared during the initial stage, we found that these instructions alone were not sufficient for training the annotators. This was primarily because of the habit of annotators to imagine and write about objects and scenes that were related to the image set, but not present in it. Frequent use of adjectives and descriptive language, neither of which were inferable given the image set, was a second major problem. We conducted multiple rounds of test annotations and live training for the annotators in order to overcome these problems.
Guidelines for labelling image sets:

1) Use simple sentences to describe the objects/places seen in the set of photos. The descriptions should be such that they are helpful for tourists visiting the location shown in the photos.

2) Each description should be 2-3 sentences long with the following structure:
   
   a) **First sentence** must describe **only the main subject** in the set of photos. 
   
   Eg: “This is an urban park.”
   “This is a hiking trail.” or
   “This is a cliff near the sea.”

   b) **Second and third sentences** should describe **all other major objects/scenes** in the photos that are of interest.
   
   Eg: (This is an urban park.) “There is a statue in the park. There is also a children’s play area.”
   (This is a hiking trail.) “The trail runs over many small hills. The trail leads to a secluded lake.”
   (This is a cliff near the sea.) “There are many birds nesting on the cliff sides. There is a hiking trail over the cliff.”

3) Try to describe as many objects/scenes as you can using simple sentences. **Do not skip any major object/scene even if it is present only in a single photo.**
   
   Eg: “This is an urban park. There is a statue with lawns and gardens around it. There is a children's play area in the park.”

4) The labeler may skip a photo set if there is no describable common theme in the photo collection, or if a 2-3 sentence description cannot be made.
Below are the names of some objects that may be seen in the photo sets. The labeler may use these names to describe the objects.

**Gazebo** - a roofed structure that offers an open view of the surrounding area, typically used for relaxation or entertainment

![Gazebo](image1.jpg) ![Gazebo](image2.jpg)

**Promenade** - a paved public walk, typically one along a waterfront

![Promenade](image3.jpg) ![Promenade](image4.jpg)

**Pier** - a platform supported on pillars or girders leading out from the shore into a body of water, used as a landing stage for boats

![Pier](image5.jpg) ![Pier](image6.jpg)
**Dock** – a dock is like a pier, but with detailed provision for boats to be parked for longer periods

**Trail** – a walking / hiking / horse riding track, usually unpaved

**Walkway** – a clearly defined path for pedestrians, usually paved

**Lawn** – an area of short mown grass

**Grassland** – an area where the natural vegetation is predominantly grass

**Waterfall**

**Cascade** - a small waterfall

Note that there will be many more objects/scenes that appear in photo sets. Please describe them using appropriate names.