# Diverse Neural Photo Album Summarization

Yunus Emre Ozkose, Bora Celikkale, Erkut Erdem and Aykut Erdem
Hacettepe University, Ankara, Turkey
e-mail: {aliozkose,ibcelikkale,erkut,aykut}@hacettepe.edu.tr

***Abstract*—In this paper, we address the problem of learning to summarize personal photo albums. That is, given a photo album, we aim to select a small set of representative images from the album so that the extracted summary captures most of the story that is being told through the images. More specifically, we extend a recently proposed recurrent neural network based framework by employing a more effective way to represent images and, more importantly, adding a diversity term to the main objective. Our diversity term is based on the idea of jointly training a discriminator network to evaluate the diversity of the selected images. This alleviates the issue of selecting near-duplicate or semantically similar images, which is the primary shortcoming of the base approach. The experimental results show that our improved model produces better or comparable summaries, providing a good balance between quality and diversity.**

Photo album summarization, deep learning, recurrent neural networks, diversity analysis

## I. Introduction

Fast development and spread of digital cameras brought rapid increase in the amount visual data. Nowadays, a person can take thousands, if not tens of thousands, photos within a single year through a camera or a smart phone. Most of the time, these photos are shared in the social media platforms such as Instagram, Facebook or Flickr, but manually selecting photos from albums and managing them is hugely time-consuming. Even if these photos are automatically organized within different albums or collections based on date or location, it makes really hard to manage these collections to obtain a visual summary of the memories grabbed in these albums. In that regard, a photo album can be interpreted a set of images containing certain events. The task of photo album summarization is simply defined as selecting a set of representative images from a photo album [1, 3, 9, 10, 13, 17, 18]. Albums generally tend to contain semantically similar or near duplicate images which cover several different events observed within an album. Hence, for summarization, the main challenge lies in understanding the album in a global manner.

Figure 1 demonstrates this process on a sample photo album created from a trip to Venice, Italy. As can be seen from the ground truth summary created by a human subject, while generating a summary, humans generally try to cover the whole story of the album with a diverse set of images. Hence, within a summary, there should be no particularly similar or near duplicate images. Of course, these characteristics of a summary requires understanding the temporal relationships among images and image semantics, which is not an easy task. Photo albums can be related to several concepts and may span over different periods of time. For example, if the central theme of the album is traveling, images can be taken over days, a week or longer periods of time. However, if it is about a wedding, generally the time span is in hours or a day. These variations in photo albums makes the summarization task difficult. For instance, traveling London can contains similar human behaviors such that tourists tend to visit same consecutive locations. On the other hand, if the theme of a photo album is wedding, it is really hard to understand the events as different cultures have different wedding ceremonies leading to a variety of different stories in the visual domain.

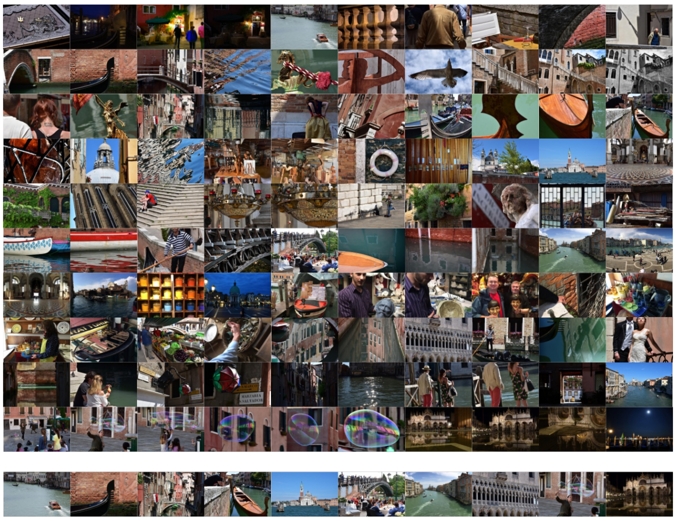

Fig. 1. Creating visual summary of a given photo album requires selecting a set of representative images that are different from each other but have a semantic temporal coherence. Here, the bottom row shows a human-generated summary of 10 images for a photo album containing 100 images about a trip to Venice, Italy.

Albums sometimes contain photos which are not related in conceptual terms. For example, consider a photo album created for a snowboarding trip. When people go to snowboarding, they visit not only the snowy mountains but also other places, certain parts of the city as well and take photos there. Hence, a summarization method needs to cover all of these events in the photo albums. Another key challenge is that summarization is generally considered as a subjective task. Hence, the evaluation of summarization methods is generally carried out by comparing the automatically created summaries against a set of summaries produced by human subjects. For example, consider the

978-1-7281-3975-3/19/$31.00 ©2019 IEEE

two photo albums shown in Fig. 2, one about snowboarding, the other about a trip to London. For each album, the first two rows shows the summaries created by different people, and the last row is a summary automatically generated by the Skipping RNN model by Sigurdsson et al. [13]. As can be seen, human generated summaries are more diverse and more inclusive whereas the Skipping RNN model fails to cover the events captured in the albums and its summaries contain quite similar images. In our work, we extend the Skipping RNN model [13] by introducing a new objective function for training that discourages including semantically similar or near duplicate images in the summaries.

Our paper is organized as follows: In section II, we first briefly discuss the related work. In section III, we provide the background and introduce our proposed framework. Then, in Section IV, we present and discuss our experimental results.

## II. Related Work

A line previous works have investigated how to best represent images for the summarization task. Li et al. [10] propose a two-stage framework. In the first stage, the album is partitioned according to first the time information and then the image content. To represent images, the authors employ color histogram as well as detected faces to better localize the important content. The method in [9] also follows a person-centric view and performs photo summarization by hierarchical clustering taking into account the color histograms computed over the face regions in computing pairwise image similarities. All these works assume that images containing persons are important for the summaries and hence totally ignore scenic images. Even if personal photo albums are full of person images, we believe that putting a special emphasis on faces puts too much constraint on the summarization task.

In the literature, a lot of effort has been devoted to eliminate redundant images to be selected for the summary. Chang et al. [3] propose a supervised approach where they combine several hand-crafted features (face, composition, clarity, deep features) to encode images and train a random forest classifier and SVM to predict the images of the best quality. Moreover, they also tested a neural network model based on a siamese architecture [2]. Bosselut et al. [1] exploit both visual and textual features to learn events from a large collection of albums. The summary is obtained by a clustering algorithm based on extracted events and learned temporal knowledge between the events. There is also a people-centric method [17] which uses mixed integer linear programming to find sub-events. Zhang et al. [18] propose a method which clusters photos with low-level features (color, texture, SIFT, etc.), employ hidden Markov model and a Gaussian mixture model to fin sub-events and find representative photos for each cluster.

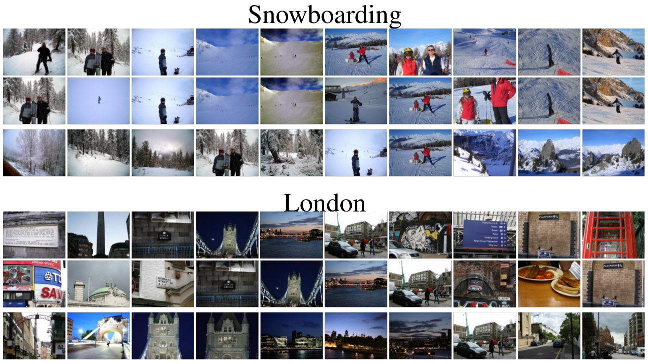

Fig. 2. Sample summaries created by humans (first two rows) and the automated summarization method Skipping-RNN [13] (bottom row) for two different photo albums. These albums are from [13].

The most similar work to ours is [13] which is, in fact, our baseline. Sigurdsson et al. [13] propose a recurrent neural network which learns a transition by skipping a lot of images in the sequences instead of consecutive transition. The model takes samples while training and update the network in an unsupervised manner. In this way, photo sequences are skipped and a common story is learned. Details are given in section III-B.

## III. Our Approach

In this section, we first provide the background by discussing the standard recurrent neural networks and then briefly review the Skipping RNN model [13]. Following that, we give the details of our summarization framework by discussing how we extend the Skipping RNN model so that it gives more diverse summaries.

### A. Recurrent Neural Networks

Recurrent neural networks (RNNs) [4] are sequence to sequence prediction models proposed to deal with sequential data. There are many different types of RNNs that consider one to many, many to many or many to one prediction task. Summarization task needs to be modelled as a many to many prediction task since we have multiple images within an album as input and multiple outputs that denote the indices of the images selected for the summary. RNNs consist of three layers, namely input, hidden and output layers, which can be modelled as follows:

$$\begin{aligned} h^t &= W_h h^{t-1} + W_x x^t \\ y^t &= W_y h^t \end{aligned} \tag{1}$$

Here, $W_x$ denotes the weights for the input layer, $W_h$ represents the weights for the hidden layer, and $W_y$ denotes the weights for the output layer. Simply, in the vanilla RNN models, input vector and previous hidden vector generate a hidden vector which is then used to produce an output vector. Hidden layers can be considered as a memory which stores the previous state information about the sequence. The main problem with RNNs is that gradients may explode or vanish, which refers to the situation that gradients tend to zero or a very large value during the back-propagation phase. There are some proposed solutions such as changing the activation function, clipping the gradients [12] or using more complex RNN architectures [6], but it is still an unsolved problem to learn long-term correlations.

### B. Skipping Recurrent Neural Networks

Skipping Recurrent Neural Network (S-RNN) [13] differs from the standard RNNs used for summarization in two different ways. First of all, instead of processing every image in an album, it basically learns to skip some images in the given input sequence to better model the long-term correlations among images as shown below:

$$\mathcal{M}^* = \underset{\mathcal{M}}{\text{argmax}} \log \sum_{\mathbf{z}_{1:N}} \left( \prod_{n} P(\mathbf{x}_{z_{n+1}} | \mathbf{x}_{\mathbf{z}_{1:n}}; M) \right) P(\mathbf{z}_{1:N}) \tag{2}$$

where $\mathcal{M}$ is the model parameters, $\mathbf{z}_{1:N}$ denotes the set of indices that represent the images selected for summarization, and $\mathbf{x}_i$ is the $i$th image in the album. Expectation maximization algorithm is used for learning phase. The recurrent neural network model weights are updated by sampling given an album in unsupervised manner. Loss function is calculated over selected key images for summary.

### C. Proposed Model

Although S-RNN [13] tries to prevent the repetitive images by learning to skip certain images, its objective function does not have any term that enforces diversity. Hence, motivated by the language model proposed in [7], we propose a new framwork that we call Combined Discriminative Model (CDM) by extending S-RNN with a new objective function that explicitly favors diverse images to be included to the summaries. In particular, we add a new discriminator to the baseline S-RNN model encoding the similarities between each image pair in the given album.

$$\begin{aligned}\mathcal{M}^* = \underset{\mathcal{M}}{\text{argmax}} \Bigg( \log \sum_{\mathbf{z}_{1:N}} \left( \prod_{n} P(\mathbf{x}_{z_{n+1}} | \mathbf{x}_{\mathbf{z}_{1:n}}; \mathcal{M}) \right) \\ + \lambda s_{rep}(\mathbf{x}_{\mathbf{z}_{1:N}}) \Bigg) P(\mathbf{z}_{1:N})\end{aligned} \tag{3}$$

where $\mathcal{M}$ is the model parameters, $s_{rep}$ represents our proposed repetition score function, and $\lambda$ is a scalar denoting the overall importance of $s_{rep}$ in the whole objective function. Introducing $s_{rep}$ helps the model to avoid similar or near duplicate images in the summaries as including these kind of images increases the objective score.

In particular, while defining $s_{rep}$, we make use of the cosine similarities between the images. We compute a score $d_i$ for every image $i$ in the summary as follows:

$$d_i = \max_{j=i-k \ldots i-1} (CosSim(y_j, y_i)) \tag{4}$$

with

$$CosSim(v_0, v_1) = \frac{v_0 \cdot v_1}{||v_0|| \cdot ||v_1||} \ , \tag{5}$$

where $k$ denote a fixed temporal window and $\mathbf{y}_i$ represents the feature vectors of an image selected by the S-RNN model. Uniqueness of an image is measured by the maximum pairwise similarity score within the specified temporal window. Our repetition score function is then defined by:

$$s_{rep}(\mathbf{y}) = \sigma(\mathbf{w}_r^T RNN_{rep}(d)) \tag{6}$$

where $RNN_{rep}(d)$ denotes the final state of a unidirection RNN model with a hidden layer size of 100 and defined over the series of similarity scores $\mathbf{d} = d_1 \ldots d_n$ and $\mathbf{w}_r$ represents a learnable vector. Especially, learning $\mathbf{w}_r$ is carried out by maximizing the following ranking log likelihood:.

$$L_{rep} = \sum_{\mathbf{y}_s \sim S-RNN(\mathbf{x})} log\ \sigma(\beta - s_{rep}(\mathbf{y}_s)) \tag{7}$$

where $\beta$ is a scalar denoting a constraint for the similarities.

S-RNN model [13] employs the output of the fc7 layer of the pre-trained Alexnet model [8] on Imagenet, which is a 4096 dimentional feature vector, to represent images. In our model, to better capture the similarities between images, we use the recently proposed R-MAC representation [15] which exploits the convolutional layers by focusing on certain image regions. First, we apply VGG16 [14] to an input image *I*, sample some regions over the last convolution layer and apply Region-of-Interest (RoI) pooling. Finally, after applying fully connected layers to the regions, we combined their outputs to produce 512-dimensional feature vectors as our image representations.

## IV. Experimental Results

### A. Implementation and Training Details

In our experiments, for fairness reasons, we conduct our analysis with R-MAC features for both S-RNN and CDM. We set the similarity constraint $\beta$ to 0.4, $\lambda$ denoting the overall importance of $s_{rep}$ to 1, and the temporals window size $k$ to 2. During training the repetition discriminator $s_{rep}$, we set the learning rate to 0.0001 and the number of iterations as 10.000 iterations and use the stochastic gradient descent (SGD) .

We carry out our summarization experiments on the CitySum dataset [5] which includes 12 different photo collections from 5 different cities, namely Amsterdam, Paris, Tokyo, New York and Venice. Each of these photo albums consists of 100 images, and each album contains 20 ground truth summaries containing 10 images which are collected by human subjects. Since this dataset is small in size, we gather a large set of photo collections from Flickr by querying the cities available in the CitySum dataset and use them to train the models for each city. The statistics of these collections are given in Table I. In our experiments, as in [13], we sample the summaries 500 times by employing both S-RNN and CDM and then take

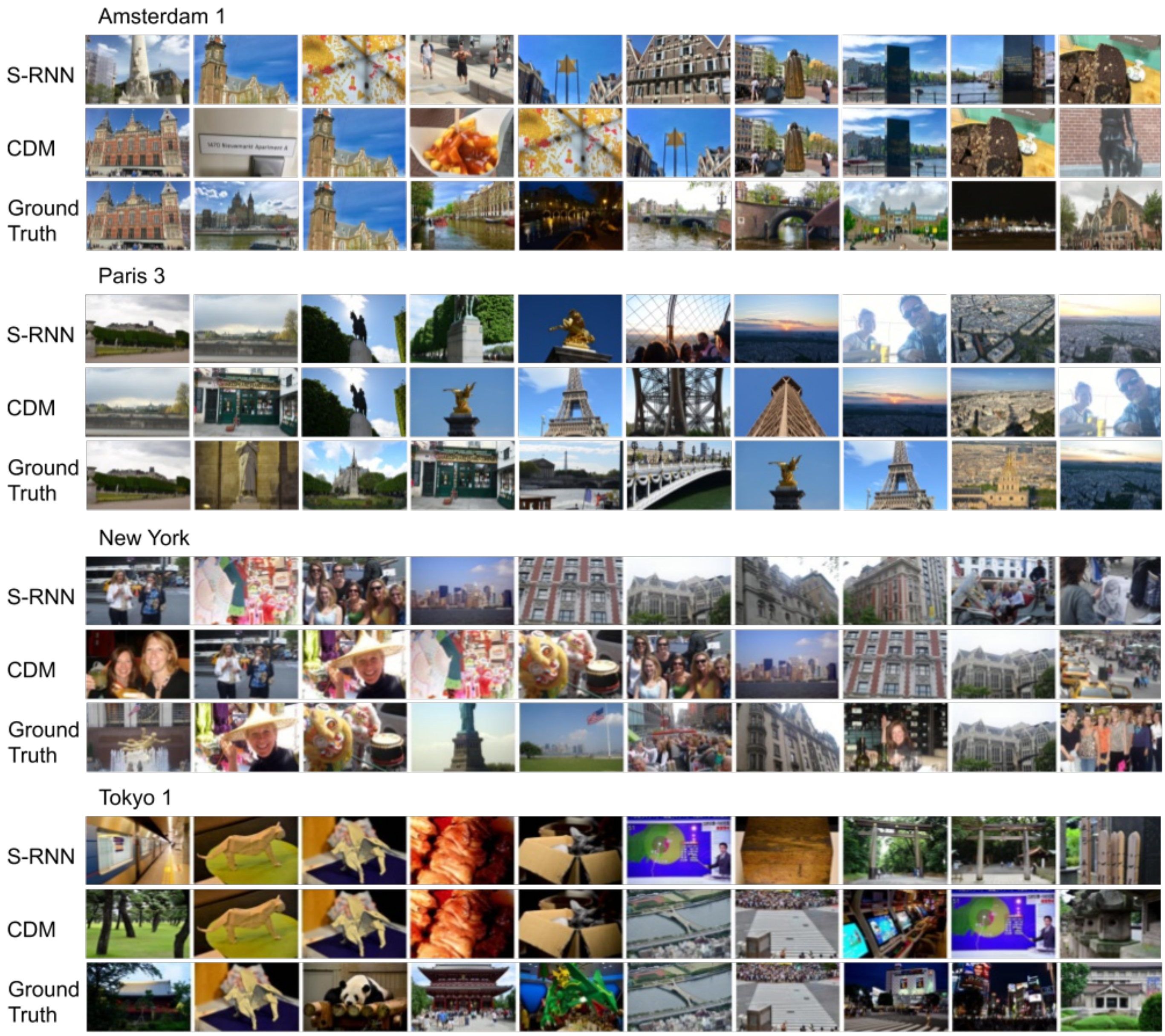

Fig. 3. Sample summaries obtained by S-RNN and CDM along with a ground truth summary created by a human. CDM provides visually more diverse summaries than S-RNN.

TABLE I
STATISTICS OF THE DATASETS USED IN TRAINING.

| Dataset | # of albums | Avg. # of images in the albums |
|---|---|---|
| Amsterdam | 100 | 99 |
| New York | 1154 | 75 |
| Paris | 728 | 90 |
| Tokyo | 932 | 83 |
| Venice | 265 | 120 |

the summary with the highest overall likelihood for each test album.

### B. *Qualitative Analysis*

In Fig. 3, we show some sample summarization results obtained with S-RNN and our proposed approach CDM for four different city albums. Although there are some common images in the summaries, S-RNN does not disfavor inclusion of visually similar images in the summaries. For instance, for the New York album there are many buildings images in the created summary. Similarly, for the Tokyo 1 album, S-RNN selects two different wooden gate images for the summary. On the other hand, the proposed CDM approach gives visually more diverse images and the number of visually similar images is limited as compared to S-RNN.

TABLE II
QUANTITATIVE ANALYSIS OF S-RNN AND CDM IN TERMS OF DIVERSITY.

| | Cosine Similarity | | LPIPS | |
|---|---|---|---|---|
| Photo Streams | S-RNN | CDM | S-RNN | CDM |
| Amsterdam 1 | **0.9936** | 0.9943 | **0.6579** | **0.6579** |
| Amsterdam 2 | **0.9941** | 0.9947 | **0.6096** | 0.5779 |
| Newyork | 0.9937 | **0.9935** | 0.6780 | **0.6984** |
| Paris 1 | **0.9923** | **0.9923** | 0.6344 | **0.6447** |
| Paris 2 | **0.9929** | 0.9937 | 0.6161 | **0.6219** |
| Paris 3 | 0.9920 | **0.9903** | **0.6590** | 0.6502 |
| Paris 4 | **0.9947** | 0.9950 | **0.6095** | 0.5827 |
| Paris 5 | 0.9947 | **0.9944** | 0.6105 | **0.6267** |
| Tokyo 1 | **0.9910** | 0.9913 | 0.7307 | **0.7344** |
| Tokyo 2 | **0.9934** | 0.9937 | **0.6873** | 0.6823 |
| Venice 1 | **0.9900** | 0.9926 | 0.6941 | **0.7096** |
| Venice 2 | **0.9909** | 0.9913 | 0.6947 | **0.6794** |

### C. Quantitative Results

We quantitatively evaluate the performance of the proposed CDM approach in two different aspects. The first aspect is the diversity of the images being selected for the album summaries. For the second aspect, we simply analyze the summarization capability of CDM in terms of quality by comparing the summaries generated by our approach against ground truth human summaries.

For our diversity experiments, we generate 50 summaries for both our approach and the baseline S-RNN model and calculate both pairwise Cosine Similarities and Learned Perceptual Image Patch (LPIPS) metric [19] of the images in the summaries in pairs. For diversity, higher LPIPS distances and lower Cosine Similarities are preferred. Hence, for each summary sample, we estimate the minimum pairwise Cosine Similarities and report the maximum of those similarity scores among the generated samples. Similary, we estimate the maximum LPIPS distances among the summary images for each summary sample, and report the minimum of those distances among the generated samples. Table II presents these results. As can be seen from this table, our analysis show that our improved CDM model produces better or comparable summaries to S-RNN in terms of diversity.

To examine the summarization performance of the proposed model, we compared the summaries by CDM and S-RNN against the summaries generated by 20 different human subjects. In that regard, we consider the commonly used F-score and V-ROUGE [16], a visual variant of the ROUGE metric[11] commonly used for text summarization. While F-score checks how much the images selected by the humans and the automatic summarization methods match within a summary, V-ROUGE metric considers visual similarities between the images in the human summaries and the images selected by the summarization methods. Table III summarizes the performances of S-RNN and CDM. As can be seen from these results, our proposed CDM method gives quantitatively better results than S-RNN in the majority of the tested photo albums in terms of F-score and V-ROUGE.

TABLE III
QUANTITATIVE ANALYSIS OF S-RNN AND CDM IN TERMS OF SUMMARIZATION PERFORMANCE.

| | F-score | | V-ROUGE | |
|---|---|---|---|---|
| Album | S-RNN | CDM | S-RNN | CDM |
| Amsterdam 1 | 0.0516 | **0.0525** | 0.3019 | **0.3030** |
| Amsterdam 2 | **0.0854** | 0.0849 | 0.4350 | **0.4385** |
| New York | **0.1118** | 0.1116 | **0.4083** | 0.4053 |
| Paris 1 | **0.0709** | **0.0709** | 0.3403 | **0.3548** |
| Paris 2 | **0.1519** | 0.1480 | **0.5010** | 0.4976 |
| Paris 3 | 0.0671 | **0.0695** | 0.3240 | **0.3279** |
| Paris 4 | **0.1635** | 0.1479 | **0.3825** | 0.3695 |
| Paris 5 | 0.0758 | **0.0783** | 0.4490 | **0.4543** |
| Tokyo 1 | 0.0747 | **0.0789** | 0.4117 | **0.4213** |
| Tokyo 2 | **0.0875** | 0.0823 | **0.4389** | 0.4302 |
| Venice 1 | 0.0679 | **0.0782** | 0.3234 | **0.3338** |
| Venice 2 | 0.0955 | **0.0956** | **0.3574** | 0.3571 |

## V. CONCLUSION

We proposed a novel photo album summarization method which is built on a novel recurrent neural networks model. In particular, we extend the S-RNN model by [13] with a combined training loss which avoids the repetitions in the generated summaries of big image collections while keeping the long-term correlations intact. Our qualitative results demonstrate that our proposed method improves the quality of summaries to a certain extent by decreasing the likelihood of selecting similar or near-duplicate images as the representative samples of a photo album. Moreover, our quantitative results show that the proposed method generates visual summaries of the photo albums which are more correlated with the ground truth summaries generated by humans as compared to the baseline method S-RNN [13].

## ACKNOWLEDGMENT

This work was supported in part by TUBA GEBIP fellowship awarded to E. Erdem and by The Scientific and Technological Research Council of Turkey (TUBITAK) Primary Subjects R&D Funding Program Award No. 116E685.

## REFERENCES

[1] Antoine Bosselut, Jianfu Chen, David Warren, Hannaneh Hajishirzi, and Yejin Choi. Learning prototypical event structure from photo albums. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1769–1779, 2016.

[2] Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. Signature verification using a" siamese" time delay neural network. In *Advances in neural information processing systems*, pages 737–744, 1994.

[3] Huiwen Chang, Fisher Yu, Jue Wang, Douglas Ashley, and Adam Finkelstein. Automatic Triage for a Photo Series.

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

[5] Goksu Erdogan. Image collection summarization with intrinsic properties. Master's thesis, Hacettepe University, Ankara, Turkey, 2018. Thesis No: 2018 YL 58362.

[6] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[7] Ari Holtzman, Jan Buys, Maxwell Forbes, Antoine Bosselut, David Golub, and Yejin Choi. Learning to write with cooperative discriminators. In *Proceedings of the Association for Computational Linguistics*, 2018.

[8] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[9] Cheng-Hung Li, Chih-Yi Chiu, Chun-Rong Huang, Chu-Song Chen, and Lee-Feng Chien. Image content clustering and summarization for photo collections. In *2006 IEEE International Conference on Multimedia and Expo*, pages 1033–1036. IEEE, 2006.

[10] Jun Li, Joo Hwee Lim, and Qi Tian. Automatic summarization for personal digital photos. In *Proc. of ICICS-PCM*, volume 3. Citeseer, 2003.

[11] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out*, 2004.

[12] Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. On the difficulty of training recurrent neural networks. In *International Conference on Machine Learning*, pages 1310–1318, 2013.

[13] Gunnar A Sigurdsson, Xinlei Chen, and Abhinav Gupta. Learning visual storylines with skipping recurrent neural networks. In *European Conference on Computer Vision*, pages 71–88. Springer, 2016.

[14] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[15] Giorgos Tolias, Ronan Sicre, and Hervé Jégou. Particular object retrieval with integral max-pooling of cnn activations. 2016.

[16] Sebastian Tschiatschek, Rishabh K Iyer, Haochen Wei, and Jeff A Bilmes. Learning mixtures of submodular functions for image collection summarization. In *Advances in neural information processing systems*, pages 1413–1421, 2014.

[17] Vassilios Vonikakis, Ramanathan Subramanian, Jonas Arnfred, and Stefan Winkler. A probabilistic approach to people-centric photo selection and sequencing. *IEEE Transactions on Multimedia*, 19(11):2609–2624, 2017.

[18] Liyan Zhang, Bradley Denney, and Juwei Lu. Subevent recognition and summarization for structured scenario photos. *Multimedia Tools and Applications*, 75(15):9295–9314, 2016.

[19] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.