SUMMARY GENERATION FOR IMAGE CORPUS USING GAN

A V Subramanyam

Indraprastha Institute of Information Technology
New Delhi

Anurag Singh, Lakshay Virmani

Netaji Subhas Institute of Technology
University of Delhi
New Delhi

ABSTRACT

In this paper, we propose a novel approach for image corpus summarization using a generative adversarial network (GAN). Our unsupervised technique can be used to automatically provide a condensed set of representatives for the given image collection. The generated summaries of data sets can be used for rapid prototyping as models can be trained using the summarized set instead of the larger original data set.The problem is challenging because a good summary must cover various aspects of an image set such as relevance, diversity and significance. The incorporation of these aspects is non-trivial because images in collections can be largely unrelated in terms of scene or context. Additionally, lack of sufficient ground truth data makes the problem very hard to solve using classical machine learning approaches. In our algorithm, we use a CNN based embedding and a score layer to compute the priority of each image towards the summary. Out network is trained in an unsupervised manners using a generator for generating the summary and a discriminator classifying between original and summary. The summaries are evaluated for diversity based on Gini Coefficient and F-Scores. We analyze the performance of our algorithm against K-Means Clustering of the features of the image set. We also provide a quantitative measurement for assessing the quality of the image summary by investigating the classification accuracy achieved when an InceptionV3 model is fine-tuned using the summary, and the original image set.

Index Terms— Gini Coefficient, Image Collection Summarization, Generative Adversarial Network, Information Visualization

1. INTRODUCTION

Image corpus summarization is an essential requirement for efficient representation, navigation and exploration. Web image collection for e-commerce, tourism and travel exploration, personal album collections, online image recommendation systems are some of the immediate applications of automatic image corpus summarization[1]. A very important application is while training machine learning algorithms. While video summarization has been well explored for efficient browsing [2] [3], image corpus summarization has received far less attention [4]. We can train machine learning models for different applications using smaller and precise data sets built as a summary of huge data sets. Summarization of data set can help train models without trading-off much on accuracy as the diversity of data will be maintained.
Summary of image corpus can be both qualitatively and quantitatively analyzed based on factor of relevance that is how relevant is a particular image to our task for which summary is being built and also the factor of diversity that all the images that are distinct are included in the summary i.e. cover all the aspects of data set and must not contain any redundancy [5]. The models for image collection summarization similar to any other automatic summarization problem can be categorized. Summarization techniques can be categorized into the following: simultaneous [5][6][7] and iterative[8].We look at all the images simultaneously or their feature vectors at same time and pick some for summary. Or an iterative model which learns about the data set after completely iterating it multiple times. While the simultaneous models are good there is no efficient way to critically analyze large data all at once after certain point[9]. Thus challenging task is to create a model that summarizes by looking at small set of images at once in multiple iterations. Collections are diverse and the order might not have any temporal sequence or correlation within. A way to inherently identify how important a image is in set of images is critical. The main contributions of this paper are:

1. We propose a novel iterative image summarization model using GAN against another baseline simultaneous model which clusters the data set using k-means and picks each center of cluster as summary.

2. We re-introduce an evaluation metric which is often used in economics called Gini Coefficient[10]. Gini Coefficient can be used to quantify how diverse a summary is. The lower the Gini Coefficient more is the diversity.

3. We analyze the diversity and relevance of summary by training a classifier on it and comparing the performance against usage of original data set.

2. RELATED WORKS

The field of image collection summarization is an important field in information retrieval, machine learning and multimedia processing. The work done by J Camero et. al. [6] gives a multi-kernel clustering approach for the image summarization task. Many of the works use clustering in on way or another [7] [5]. Some other approaches involve graph methods[11] [12],similarity methods[13] or neural network based methods[8] such as self organizing maps. The authors in [5] consider scene summarization which deals with problem of concisely depicting a scene that is creating a visual summary for a given interest point.Images in a summary must not be totally identical to each other.This concept is referred as diversity or dispersion the property is also known as orthogonality. It can also be noted that authors in [5] give a concept of "likelihood" where image representing a set of images in summary must have similarity to that set of images. But it can be noted that the concept of diversity is not only for majority of images in original data set but also applies to minority.With the constraints of size and other parameters diversity and relevance must be maximized such that all the aspects i.e. both minority and majority must be included in summary. In 2013 Yang et. al. [1] formulate image summarization as an optimization problem. The authors apply a dictionary learning approach based of SIFT-Bag of Words model for creating the summary. In their NIPS 2014 paper Sebastian et. al. [4], pose a detailed analysis and try to solve problem of image summarization using sub modular functions.

In addition the authors in [4] also propose a novel evaluation metric which they name V-ROUGE based on recall inspired by ROUGE a evaluation metric extensively used in document summarization community. The precision and recall metrics define quantitatively how good a summary only when they are provided with user annotated summary. This limits the scope to data sets where not much meta data is available. Secondly this also adds a issue to annotate the ground truth with relevance to a particular task for which summary is desired. As it is possible that which change of task the images needed in summary may change.

3. IMAGE COLLECTION SUMMARIZATION

Out network takes an input in terms of CNN feature embedding of images followed by a score layer. The fused output of score layer with CNN feature vectors are used input to GAN.

3.1. Problem Formulation

Summarization can be approached in two different ways as a subset selection problem or as an optimization problem. Given a collection of n images $X = (X_1, X_2, ..., X_n)$ we aim to find a subset The subset S such that $S \subset X$ and $|S| < n$, while preserving the relevance and diversity.

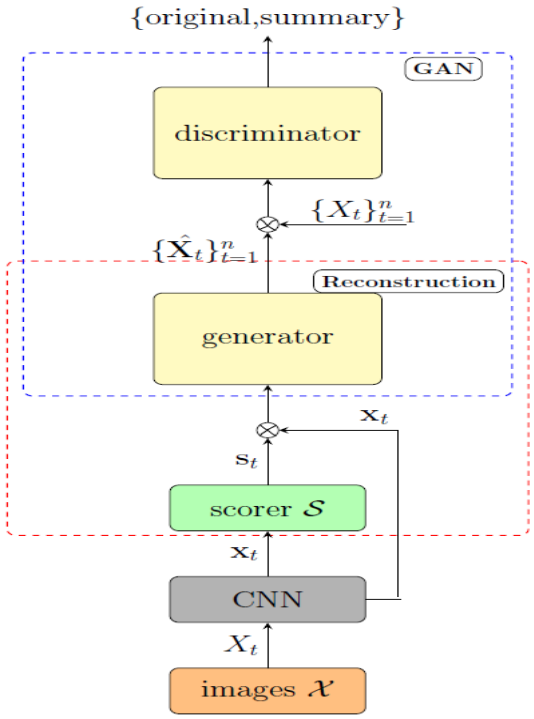

Fig. 1. Main components of our approach: The scorer ANN (sANN) gives score s for each image which defines its importance for the summary. Then each deep feature vector is weighted by its score and forwarded to generator called as gCNN for reconstructing the image collection $\hat{x}$. The discriminator (cCNN) classifies $\hat{x}$ as original or summary class.And they both form the generative adversarial network (GAN).

3.2. Learning Framework

In our algorithm, we first extract features of the images $X = \{X_1, X_2, ..., X_n\}$ using inception v3 [14]. Let these features be $x = \{x_t : t = 1...n\}$.These features are fed as an input to a score layer score ANN (sANN). These scores represent the relative importance of the image being present in the summary. Let $s = \{s_t : s_t \in [0, 1]t = 1...n\}$ be the scores, where $s_t \in [0, 1]$.The features x_t are weighted using these scores only. This then acts as a input to the generator which then reconstructs the image collection as a sequence of images. $\hat{X} = \{\hat{X}_1, \hat{X}_2, ..., \hat{X}_n\}$.
The discriminator in the GAN is aimed to classify images as two distinct classes. Thus distinguish between images from X and $\hat{X}$ as 'Original' and 'Summary'.The discriminator(cCNN) can be thought as a estimator of distance between X and $\hat{X}$ and assigning different class labels to both of them if they are distinguishable. Therefore, it can be concluded that discriminator acts as a means to represent the

error between the original image collection and reconstructed collection from the summary.GAN in the model is Similar to generative adversarial networks presented in DC-GANs and face-GANs[15].Generator and discriminator are trained adversarially until the discriminator is not able to discriminate between the re-collection and original collection.

3.3. Training the model

We discuss the different loss functions and training part of the algorithm in this section. The parameters of model are w_s, w_g, w_d for the scorer, generator and discriminator. The training of our model is defined by following losses Loss of GAN $\mathcal{L}_{GAN}$. Reconstruction loss $\mathcal{L}_{reconstruct}$. And regularization loss $\mathcal{L}_{sparsity}$. Similar to usual training GAN models in adversarial manner.The objective is iteratively achieved by:

1. for learning $\{w_s\}$, minimize $\mathcal{L}_{sparisity} + \mathcal{L}_{reconstruct}$
2. for learning $\{w_g\}$, minimize $\mathcal{L}_{GAN} + \mathcal{L}_{reconstruct}$
3. for learning $\{w_d\}$, maximize $\mathcal{L}_{GAN}$

3.4. Reconstruction Loss $\mathcal{L}_{reconstruct}$

$\mathcal{L}_{reconstruct}$ is used to make a summary that captures all relevant frames. If original set of images can be reconstructed using the the deep feature vectors weighted by summary scores. Then summary can be considered to have contained all relevant frames. $\mathcal{L}_{reconstuct} = \|\sum_{t=1}^{n}(\mathbf{x_t} - \hat{\mathbf{x}}_\mathbf{t})\|$ Where x_t is image in dataset, $\hat{x}_t$ is reconstructed from the generator and n is number of images in dataset.

3.5. Loss of GAN $\mathcal{L}_{GAN}$

Similar to [16] we train the classifier i.e. discriminator such that it is able to distinguish between the 'original' x and 'summary' $\hat{x}$. The $\mathcal{L}_{GAN}$ is thus defined as:

$$\mathcal{L}_{GAN} = log(cCNN(x)) + log(1 - cCNN(\hat{x})) \quad (1)$$

where $cCNN(.)$ is the soft-max output of discriminator.

3.6. Sparsity loss $\mathcal{L}_{sparsity}$

The sparsity loss acts as the regularization loss for the model, which regularizes the number of images that form the summary. Regularization is important to ensure non redundancy in the data set so that summary length is minimal.

$$\mathcal{L}_{sparsity} = \left\| \frac{1}{n} \sum_{1}^{n} s_t - \sigma \right\|_2 \quad (2)$$

Where σ is a hyper-parameter to control the percentage of images in summary.

Algorithm 1 Training the model

function UPDATE PARAMS ▷ where input is the feature vector secquence and output is learned parameters w_s, w_d, wc
  for max number of iterations **do**
    $X \leftarrow MiniBatchOfImages$
    $x \leftarrow CNN(X)$
    $S \leftarrow sANN(x)$ ▷ select frames
    $E \leftarrow x * S$ ▷ weight vectors by scores
    $\hat{X} \leftarrow gCNN(E)$ ▷ Reconstruction
    $\{w_s\} -= \nabla(\mathcal{L}_{reconstruct} + \mathcal{L}_{sparsity})$
    $\{w_g\} -= \nabla(\mathcal{L}_{reconstruct} + \mathcal{L}_{GAN})$
    $\{w_d\} += \nabla(\mathcal{L}_{GAN}) \% MaximizationUpdate$

4. RESULTS

The approach is evaluated using following datasets: CIFAR-10 [17], CIFAR-100 [17] Animals with attributes 2 (AwA2) [18], VOC2012 [19] and diversity 2016 [20]. CIFAR-10 consists of 60,000 32X32 tiny images belonging to 10 classes with same images per class.There are 50,000 training and 10,000 test images. CIFAR-100 consists of similar 60,000 32x32 tiny images with 600 images per class. The classes are divided into 20 super-classes with 5 classes per super-class. AwA2 is another data set used for classification purposes. It contains 37322 images of 50 animal classes. Visual Object Classes (VOC 2012) is another image classification data set with 20 classes and 11,530 images. While diversity 2016 contains the images with corresponding ground truth images for task of diversity in image retrieval. Images are ranked according to their importance within a class in ground truth annotations. There are 20821 images of multiple classes with each class containing 300 images. The classes correspond to events such as balloon festival, Buckingham guard change, Diwali or sports like surfing etc.

4.1. Evaluation Metrics

Regarding evaluation metrics there have been multiple attempts to understand summaries both quantitatively and qualitatively.Still there exists a need for gold standard both in terms of data set with annotated summaries and evaluation of summaries generated for data sets with only meta data being image labels. Like previous works in video summarization [21] [3] where key frame annotations are given, Precision and Recall can be used for evaluation of summaries of data sets with ground truths. precision is ratio of number of correct classifications to summary length and recall is ratio of number of correct classifications to size of ground truth. The F-score is harmonic mean of two. To measure the diversity of our summary on data sets with only meta data available as image labels we use gini-index[10]. A metric often used in economics to define diversity of income levels in a country. In

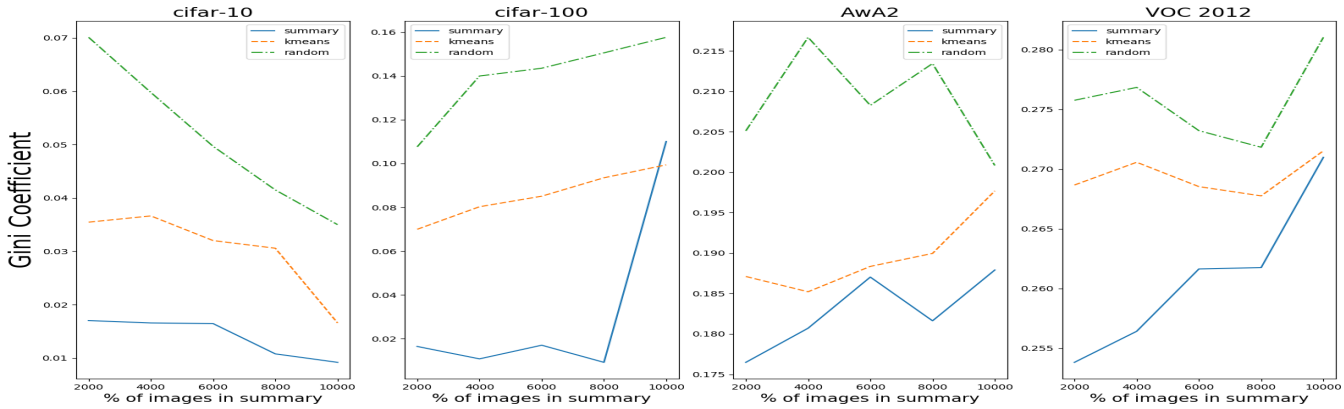

Fig. 2. The gini coefficient for different number of images in summary of multiple datasets. With images selected at random from data set, selected using Kmeans clustering and iamegs selected in summary from the model

σ	F-score	σ	F=score
0.1	88.1	0.6	83.8
0.2	66.2	0.7	60.6
0.3	80.2	0.8	78.6
0.4	92.9	0.9	91.4
0.5	x	1	x

Table 1. Table to test captions and labels

Accuracy				
% of Data set used	100%	10%	30%	50%
CIFAR10	88.1%	78%	83.8%	x%
CIFAR100	66.2%	55.4%	60.6%	x%
VOC2012	80.2%	76.3%	78.6%	x%
AwA2	92.9%	89.5%	91.4%	x%

Table 2. Table to test captions and labels

the F-scores are plotted vs the sparsity loss hyper-parameter σ and they peak out in range $0.3 < \sigma < 0.5$. The scores drop sharply outside this range as σ tends to 0 or 1. Now for the data sets only with labels as meta data we try to plot gini-indexes as a comparison between randomly picked images from data set, images picked corresponding to centers of K-means feature clusters and gini index for images from summary are plotted in Fig:2 for different data sets. The summary is more diverse throughout the number of images present in summary with multiple data sets as lower the value of gini index more is diversity. In order to evaluate whether the summary is a good representation of original, we perform the following experiment. We first fine-tune an inception v3 model on original datasets and compute the accuracy. We report these results in Table 2. Since our goal is to test the goodness of summary, we only run the model for few epochs attaining a decent accuracy, though stat of art accuracy may be achieved by using more epochs. Further, we repeat the experiment with summary and summary with data augmentation. We observe that the accuracy achieved using summary itself is good when compared to its original counterpart. In addition, with data augmentation, the accuracy boost up and is close to original case. These results are of great significance because the training using summary uses only 10% of original data, while the trade-off in accuracy is reasonable. Thus it is evident that the summary captures most of the aspects of the dataset.

5. CONCLUSION

In this work, we propose an unsupervised model to summarize a large collection of images. From the original image set, our model selects the most diverse images in order to create a smaller, more precise summary set. In order to create summary, our model makes use of a scoring layer and fusion of CNN feature vectors with scores as input to GAN. We train the model using two different losses namely sparsity and GAN loss. The evaluation in terms of F-score show the efficiency of our algorithm. To measure diversity, we use Gini index and show the high diversity in the generated summary. In addition, we show that the classification results attained by training a deep network on summary only and on original dataset are close and show similar trend. Thus, our model can also be used for a quick analysis of various models without needing to train on entire dataset. In case the labels are not available, our technique can be used to summarize and retains fraction of data, which can be relatively convenient to annotate. Further, one can perform different tasks on this data before scaling up the model as well other processing on the original data.

6. REFERENCES

[1] Chunlei Yang, Jialie Shen, Jinye Peng, and Jianping Fan, "Image collection summarization via dictionary

learning for sparse representation," *Pattern Recognition*, vol. 46, no. 3, pp. 948–961, 2013.

[2] Aditya Khosla, Raffay Hamid, Chih-Jen Lin, and Neel Sundaresan, "Large-scale video summarization using web-image priors," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 2698–2705.

[3] Behrooz Mahasseni, Michael Lam, and Sinisa Todorovic, "Unsupervised video summarization with adversarial lstm networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[4] 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*, 2014, pp. 1413–1421.

[5] Ian Simon, Noah Snavely, and Steven M Seitz, "Scene summarization for online image collections," in *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*. IEEE, 2007, pp. 1–8.

[6] Jorge E Camargo and Fabio A González, "A multi-class kernel alignment method for image collection summarization," in *Iberoamerican Congress on Pattern Recognition*. Springer, 2009, pp. 545–552.

[7] Daniela Stan and Ishwar K Sethi, "eid: a system for exploration of image databases," *Information processing & management*, vol. 39, no. 3, pp. 335–361, 2003.

[8] Da Deng, "Content-based image collection summarization and comparison using self-organizing maps," *Pattern Recognition*, vol. 40, no. 2, pp. 718–727, 2007.

[9] CL Philip Chen and Chun-Yang Zhang, "Data-intensive applications, challenges, techniques and technologies: A survey on big data," *Information Sciences*, vol. 275, pp. 314–347, 2014.

[10] Corrado Gini, "Variabilità e mutabilità," *Reprinted in Memorie di metodologica statistica (Ed. Pizetti E, Salvemini, T). Rome: Libreria Eredi Virgilio Veschi*, 1912.

[11] Deng Cai, Xiaofei He, Zhiwei Li, Wei-Ying Ma, and Ji-Rong Wen, "Hierarchical clustering of www image search results using visual, textual and link information," in *Proceedings of the 12th annual ACM international conference on Multimedia*. ACM, 2004, pp. 952–959.

[12] Bin Gao, Tie-Yan Liu, Tao Qin, Xin Zheng, Qian-Sheng Cheng, and Wei-Ying Ma, "Web image clustering by consistent utilization of visual features and surrounding texts," in *Proceedings of the 13th annual ACM international conference on Multimedia*. ACM, 2005, pp. 112–121.

[13] Jau-Yuen Chen, Charles A Bouman, and John C Dalton, "Hierarchical browsing and search of large image databases," *IEEE transactions on Image Processing*, vol. 9, no. 3, pp. 442–455, 2000.

[14] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna, "Rethinking the inception architecture for computer vision," *CoRR*, vol. abs/1512.00567, 2015.

[15] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[16] Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, Hugo Larochelle, and Ole Winther, "Autoencoding beyond pixels using a learned similarity metric," *arXiv preprint arXiv:1512.09300*, 2015.

[17] Alex Krizhevsky and Geoffrey Hinton, "Learning multiple layers of features from tiny images," 2009.

[18] Yongqin Xian, Christoph H. Lampert, Bernt Schiele, and Zeynep Akata, "Zero-shot learning - A comprehensive evaluation of the good, the bad and the ugly," *CoRR*, vol. abs/1707.00600, 2017.

[19] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, June 2010.

[20] Bogdan Ionescu, Alexandru Lucian Gînscă, Bogdan Boteanu, Mihai Lupu, Adrian Popescu, and Henning Müller, "Div150multi: A social image retrieval result diversification dataset with multi-topic queries," in *Proceedings of the 7th International Conference on Multimedia Systems*, New York, NY, USA, 2016, MMSys '16, pp. 46:1–46:6, ACM.

[21] Yale Song, Jordi Vallmitjana, Amanda Stent, and Alejandro Jaimes, "Tvsum: Summarizing web videos using titles," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 5179–5187.