SumDL: Unsupervised Summarization of Image Collections

Anurag Singh[a], Deepak Kumar Sharma[b], Joel Rodrigues[c]

[a] *Department of Electrical Engineering, Indian Institute of Science, Bangalore*
[b] *Division of Information Technology, Netaji Subhas University of Technology, New Delhi*
[c] *UFPI - Federal University of Piaui*

Abstract

Image corpus summarization has remained a long standing challenge in the area of computer vision research. Unlike text, event or video there is no strong temporal relationship between the set of images in a dataset or even between images in a class. In fact, images in the data set that needs to be summarized need not be temporally co-related at all and are sometimes chosen with different background settings on purpose so that classifiers do not depend upon temporal data for classification. This subtlety results in significant differences regarding how image summarization must be approached. Recent developments in deep learning based architectures suggest end to end architecture for solving classification tasks and have been inspiration for work in video summarization[1][2][3]. In the proposed work our aim is to explore an alternative approach to summarization task. Using end to end model for summarization is equivalently thinking of summarization as classification task. Which is shown to have performed poorly in cases where the underlying idea is to find relationship between data points rather such as face recognition, re-identification. As compared to classification and detection, where objective is to identify relation between data and its class labels. To build a general model for image corpus summarization, model consists of CNN and (DC-GAN) [4] followed by K-means clustering. We also test our model efficacy by means of rigorous experiments both qualitatively and quantitatively. Paper also introduces novel qualitative method named ***tile visualization***, to judge the efficacy of model whereby breaking an image in set of small blocks and then using it as an input to the model.

Keywords: Image Summarization, Auto Encoders, Generative Adversarial Networks, k-Means Clustering

1. Introduction

Considering the rate at which data has been generated and rise in availability of data. It is also very much evident that there has been rise in its analysis. Organizations and companies wish to gather knowledge from data generated in various modes. In other words, multi-modal extraction of knowledge from sources such as text, images and videos. Images tend to be superior form of perception not only in anecdote but also in real world. As internet penetration continues to increase data being generated in form of images is expected to see significant rise in volume. In such cases there will be need to extensively analyze images and perform multiple tasks on them. Even the state of art algorithms cannot be trained on every image available as it would be computationally intractable. Therefore, summarization of such large corpus of data is going to be an inevitable challenge in the future. The key challenges are to rigorously define different attributes of summary. How to identify redundancy and ensure that all the parts of source data find a representation in the summarized data. Also, how to qualitatively and quantitatively evaluate a summary i.e distinguish a good and bad summary. 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

Email addresses: anurags.it@nsit.net.in (Anurag Singh), dk.sharma1982@yahoo.com (Deepak Kumar Sharma), joeljr@ieee.org (Joel Rodrigues)

[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 which helps in identification of global relationship among the data points. An iterative model is one which learns about the data set after completely iterating it multiple times. In other words, since model is never exposed to complete data at once, it has to infer the global relationships by the means of local discrimination. While the simultaneous models are good there is no efficient way to critically analyze large data simultaneously after a certain point[9]. Thus, challenging task is to create a model that summarizes by looking at a small set of images in multiple iterations. Collections are diverse and the order might not have any temporal sequence or correlation within the images. A way to inherently identify how important a image is in set of images is critical. Immediate applications of automatic image corpus summarization[10] can be to summarize datasets and collection of images in settings where there is not enough communication band width or enough time to review the whole set of generated images manually. Recent applications also include finding the representative set of best images among a burst of images captured during a event. It is also a sub-domain of event summarization . A very important application is while training machine learning algorithms. While video summarization has been well explored for efficient browsing [11] [1], image corpus summarization has received far less attention [12]. 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. Iterative Summarization for images is a challenging task as at any point during the particular iteration even if some sort of temporal data is kept by the model it is not much likely that it can preserve any meaningful information for the model to exploit, while it makes the decision to keep or reject the current image from the summary. Therefore, in this paper we try to explore the image summarization on the idea of looking at all the features at once. This although divides our model into two major components rather than being ended to end trained but it gives us added leverage to exploit and select images based on the representation of their feature vector maps. Also, it helps us consider all the images in one go so the information from all the images is utilized in making the choice to keep or discard the image from the representative set. Therefore, the main contribution of the paper is:

- To propose a novel simultaneous approach based image corpus summarization model.
- Proposed model is unsupervised and based on a GANs to extract the features from the network and then cluster those features to find out the summary by picking up images closest to the centroids of the clusters.
- To propose a novel method of tile visualization for the task of image corpus summarization, where the image is broken into small patches which are then passed through the model to get the key patches that can summarize the whole image.

The rest of this paper is structured as follows: next, Section 2 presents a comprehensive review of the approaches found in the bibliography related to image collection summarization in similar scopes. Then, Section 3 details the methodology considered in the study. Which includes the architecture of the model, types of losses used and algorithm to train the network. Datasets used for evaluation along with the evaluation metrics are reported in Section 4, which is also devoted to analyse the obtained results on different tasks and discusses about the underlying relevance diversity trade off. Visualization results are discussed in Section 5 which show the efficacy of the algorithm. Section 6 discusses about motivation and efficacy of task specific loss. Finally, conclusions and future lines of work are presented in Section 7.

2. Related Works

Summarization in case for videos has received much more recent attention and is most closely related to image summarization than text or event summarization. LSTM based models for supervised and unsupervised learning are being used to capture temporal relationships between video frames. In [1] the authors proposed a deep learning based architecture that can help to summarize different video lengths training in a unsupervised manner. They also propose the idea of using a encoder decoder

classifier based network for the task of video summarization. The network consists of a auto encoder and a GAN which share a network component. The decoder of the auto encoder also acts as the generator of the GAN. Overall this combination allows the network to have a loss that aids in gradient flow to a fully connect multi-layer perceptron(MLP) feed forward scorer which scores the confidence of the membership of the frame in summary. The input to the scorer are the features extracted from Inception V3 model. Also deep learning based models for video summarization [2] using long short term memory and their variants with application of reinforcement learning [3] show how temporal sequences can be modelled in form of states, not only with respect to time but also within the framework of reinforcement learning by making diversity representative awards. 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[13] [14],similarity methods[15] 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. [10] 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. [12], pose a detailed analysis and try to solve problem of image summarization using sub modular functions.

In addition the authors in [12] 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. Recently deep learning based approach has been experimented. In one of the seminal works around image summarization done by Jain et. al in [16], importance of intent for the task was highlighted. Authors in [16] describe about the effective image summarization and how intent and coverage play an important role in it. In many of the recent deep learning architecture developments, even in case of video summarization tend to ignore its relevance. In [17] Subramanyam et. al propose a novel end to end deep learning based architecture for iterative summarization of image corpus. They also proposed new techniques for evaluation of summaries quantitatively using Gini Coefficient and the idea of classification accuracy to test the information capturing capabilities of summary along with Reconstruction Error. They also proposed a brave new idea of considering task specific loss in training of deep learning architectures. The images in summary must be chosen keep in mind the task for which summary is intended. For example a summary needed to train a cat classifier should keep that bias in mind to give more weight to images of cats so that all the hard examples needed to train the classifier are present in the summary. In this paper we also take inspiration from [17] to use a task specific loss in our objective function. 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. Therefore a need for standard to compare different algorithms and approaches on particular dataset and fixed set of evaluation metrics needs to be established.

3. Methodology

Out network comprises of two main components: a Deep learning network and clustering. The features are extracted from the images using an encoder CNN (eCNN) and those features are then passed to decoder CNN(dCNN) which also acts a generator for the GAN. In practice, we use inception V3[18] pre-trained on Image-Net[19] for encoding the feature vectors from CNN and DC-GAN[4] for implementation of GANs. These feature embedding of images are then clustered using k means where number of clusters that is the $k = \sigma n$. Where

the n is the number of images present in the whole dataset and the σ is the fraction images to selected in the summary. The image corresponding to the feature embedding nearest to the centroid of the cluster is selected as the representative image to that cluster to be selected in the summary.

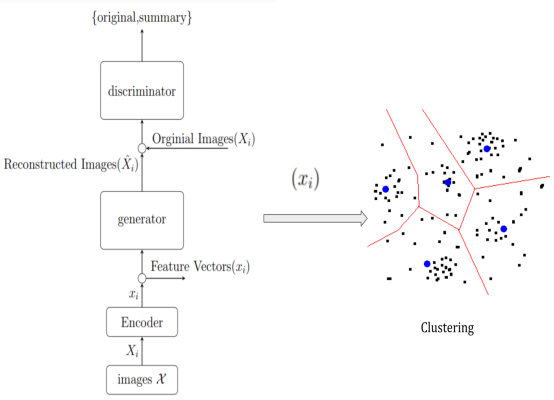


Figure 1: Main components of our approach: The feature embedding come from the encoder (eCNN). Then each deep feature vector is de-convoluted by the decoder to produce a close estimation of the original image.The reconstructed the image collection is $\hat{x}$. The discriminator (cCNN) classifies $\hat{x}$ as âoriginalâ or âsummaryâ class. And decoder and classifier both form the generative adversarial network (GAN). Thus using this model we ensure that most of the important aspects of the image get embedded into the feature vectors. So as when clustering happens it is easier to identify and group similar images based on their feature vector representation and select a representative image for the summary

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 S such that $S \subset X$ and $|S| < n$, while preserving relevance and diversity.

3.2. Learning Framework

In our algorithm, we first extract features of the images $X = \{X_1, X_2, ..., X_n\}$ using inception v3 [18]. Let these features be $x = \{x_t : t = 1...n\}$. These feature vectors are generated after 100 epochs. Features are then clustered using the k-means clustering and then the images corresponding to the feature vectors nearest to the centroids of the clusters are picked as summary. All the features acting as in latent representation of image 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'. Where 'Original' stands for the ground truth and 'Summary' for the reconstructed images. 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[4].Generator and discriminator are trained in adversarial fashion until discriminator is not able to discriminate between the generated collection and original collection. A important question to be addressed is that why are the images with feature vectors nearest to centroids being picked as a part of summary. Without loss of generality, while summarizing the dataset in general we do not know what is the perfect length of the summary that should be kept which maximizes the diversity and has least redundancy. So say for a given $k = \sigma \times n$ we pick the image nearest to centroid. For a cluster, the average distance of the centroid from every other image is minimum. Therefore, centroid serves as the best approximation for every image. Hence, we pick the image nearest to centroid as the representative image of that cluster. Here k is a hyper parameter that needs to be searched for by observing the average distance of elements in the cluster to their centroids [20][21].

3.3. Reconstruction Loss $\mathcal{L}_{recons}$

$\mathcal{L}_{recons}$ is used to make a summary that captures all relevant frames.It has been used in many of previous works in both image and video summarization [2][1][17]. If original all set of images can be reconstructed using the their deep features respectively it would mean that the network has been able to reduce the dimensionality of the problem without sacrificing much on the information present in the image. Then summary can be considered to have been generated on a tractable set of feature vector representations rather than images themselves. While keeping the encoded relevant information and making the clustering robust to negative examples where a mere pixel shift between a pair of

similar images would have landed them far in hyper-dimension space while clustering based on their euclidean distances. $\mathcal{L}_{reconstuct} = \|\sum_{t=1}^{n}(\mathbf{X_t} - \hat{\mathbf{X}}_\mathbf{t})\|$ Where X_t is image in dataset, $\hat{X}_t$ is reconstructed from the generator and n is number of images in dataset.

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

Similar to [22] 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.

3.5. Task Specific Loss $\mathcal{L}_{task}$

In [17], authors come up with a new approach for the task specific summarization and propose a method to account for the end goal for which the summary is going to be used. In this paper we also use the task specific loss implemented for the classification task.

$$\mathcal{L}_{task} = \frac{\mathcal{L}_{pre-trained}(\mathcal{X})}{\beta} \quad (2)$$

Here the standard classification task has been taken. So the images are passed from a trained network for classification task which also shares its weights with the encoder (eCNN) being the same network in case of our implementation (Inception V3)[18]. Therefore in the particular implementation of the task specific loss the eCNN part acts as a multi-task network that is both encoding of the image and its classification[23]. Multi-task learning has shown better results for training in general[24] and also for esoteric tasks such as person re-identification[25]. It is also to note that for a particular task the β is likely to control the degree to which the outliers are binned together with the inliers. Training for a particular task will ensure that the images that are labelled as belonging to same class have similar features. As mentioned in the earlier discussion, a suitable value of k can be chosen to further tune the number of outliers present in the summary.

3.6. Similarity Loss $\mathcal{L}_{sim}$

Taking inspiration from triplet loss, we introduce similarity loss $\mathcal{L}_{sim}$ in training of our feature vector encoder to ensure a more discriminatory embedding. Also, it helps in forming a feedback to the current K-means and deep learning model architecture which would otherwise be working in silos. As a feedback to the deep learning model similarity loss helps in making more discriminative features that cluster even better. Eventually, it will ensure that the model groups images with same information in the feature space together. This will ensure that the sampling of images nearest to centroid will have least reconstruction loss.

$$\mathcal{L}_{sim} = \sum_{(i,j)\in\mathcal{C}, i\neq j} ||x_i - x_j||_2^2 - \sum_{(i,k)\notin\mathcal{C}} ||x_i - x_k||_2^2 \quad (3)$$

We define $\mathcal{C} = \bigcup_{i=0}^{k=\sigma n} A_i \times A_i$ where A_i is the set of all data points in cluster i.

3.7. Training the model

We discuss the different loss functions and training part of the algorithm in this section. The parameters of model are w_e, w_d, w_c for the encoder, decoder i.e. generator and classifier i.e. discriminator. The training of our model is defined by following losses Loss of GAN $\mathcal{L}_{GAN}$. Reconstruction loss $\mathcal{L}_{recons}$. Similar to usual training GAN models in adversarial manner.The objective is iteratively achieved by:

1. for learning $\{w_e\}$, minimize $\mathcal{L}_{recons} + \mathcal{L}_{sim} + \mathcal{L}_{task}$
2. for learning $\{w_d\}$, minimize $\mathcal{L}_{GAN} + \mathcal{L}_{recons}$
3. for learning $\{w_c\}$, maximize $\mathcal{L}_{GAN}$

4. Results

4.1. Data set

The approach is evaluated using following datasets: CIFAR-10 [26], CIFAR-100 [26] Animals with attributes 2 (AwA2) [27], VOC2012 [28] and diversity 2016 [29]. 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

Algorithm 1 Training the model

function UPDATE PARAMS ▷ where input is the feature vector sequence and output is learned parameters w_e, w_d, wc
  for max number of iterations **do**
    $X \leftarrow$ Mini Batch Of Images
    $x \leftarrow eCNN(X)$ ▷ Encoding of the image into deep feature vector
    $\hat{X} \leftarrow dCNN(x)$ ▷ Reconstruction
    $\{w_e\} = \nabla(\mathcal{L}_{recons} + \mathcal{L}_{task} + \mathcal{L}_{sim})$
    $\{w_g\} - = \nabla(\mathcal{L}_{recons} + \mathcal{L}_{GAN})$
    $\{w_d\} + = \nabla(\mathcal{L}_{GAN})$ ▷ Maximization Update

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.

4.2. 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 [30] [1] where key frame annotations are given.

4.2.1. Precision and Recall

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. Here, Pr refers to precision and Re refers to recall. Let, N_s be set of images in the summary, for our model $|N_s| = \sigma n$ and N_gt refer to the set of annotated images present in the ground truth. Then precision and recall can be calculated as follows:

$$Pr = \frac{N_s \cap N_{gt}}{N_s} \tag{4}$$

$$Re = \frac{N_s \cap N_{gt}}{N_{gt}} \tag{5}$$

$$F - score = \frac{2}{\frac{1}{Pr} + \frac{1}{Re}} \tag{6}$$

In table:1 we try to compare the f-scores for different values of σ to find out how the efficacy of our model varies with respect to the length of the summary. We also try to compare our image summarization model against the two state of art video summarization techniques run on diversity 2016 dataset. We give the comparison for the two available techniques based on their open source implementation[1][3]. The three rows for each method correspond to the precision, recall and f-scores for different summary lengths given particular method. It can be observed that when methods made for video summarization are transferred to image summarization they perform relatively poorer on F-scores consistently across all the values of σ. This is partly because a lot of temporal sequence exists in videos. Video summarization architectures are designed to learn and exploit such sequences. Whereas, in the case of image datasets when no such temporal sequence may exist it is very likely that network might learn or fine tune itself on absurd parameters which can in turn hurt its performance. In table where F-scores are reported vs the sparsity loss hyper-parameter σ. They peak out in range $0.3 < \sigma < 0.5$. The scores drop sharply outside this range as σ tends to 0 or 1. This because precision value goes down with σ going from 0 to 1 because the size of summary increases and the recall goes up for same reasons.

4.2.2. Gini Coefficients

There is need for novel methods for evaluation for image corpus summarization were used when data without ground truth is summarized. To measure the diversity of our summary on data sets with only meta data available as image labels we use gini-index[31]. A metric often used in economics to define diversity of income levels in a country. In [17], authors propose use of gini coefficient for evaluation of diversity in the summarization tasks by means of understanding the representation of elements from each class in the constituent summary, which is also

Method/σ	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
Proposed	8.1	13.8	27.3	32.1	30.4	26.3	22.2	18.1	15.0	10.2
[1]	4.1	9.8	20.3	26.1	24.4	20.3	12.2	10.1	10.0	8.2
[3]	6.8	11.4	22.5	30.7	27.4	21.9	18.7	13.1	11.9	8.8

Table 1: F-Scores vs σ for the diversity 2016 dataset

referred to as completeness of the summary or its coverage capacity. Gini index is defined using a differentiable cumulative distribution function $f(x)$ which has mean ρ, and is zero for all negative values of x,

$$Gini = 1 - \frac{1}{\rho}\int_0^\infty (1 - f(x))^2 dx \qquad (7)$$

In practice, we compute an approximation to the given integral using the trapezoid rule for calculation of Gini Coefficients.

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 image clusters and gini index for images from summary are plotted in Fig:2 for different data sets. There is observed difference that our model is more diverse and shows lower gini index than the summary from clustering images directly. 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.

4.2.3. Classification Accuracy

In order to evaluate whether the summary is a good representation of original, we perform the following experiment. We first fine-tune on original datasets an inception v3 model which is pre-trained on imagenet and compute the accuracy. The results are reported 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 do not use any image augmentation techniques in the experiment with summary even for cases where summary size is 10%. We observe that the accuracy achieved using summary itself is good when compared to its original counterpart. In addition, with data augmentation, the accuracy may boost up and be significantly close to original case. Since our goal is to be able to essentially differentiate in the efficacy of summaries absolute values of accuracy do not hold much significance. However, it can be concluded that the training using summary uses only 10% of original data, while the trade-off in accuracy is reasonable it can be relevant application of summarization to use summaries to fine tune models/prototype and test deep learning architectures.

4.3. Relevance and Diversity Trade-off

In table 3, we show the number of images of each class in the VOC2012 dataset, and their respective ratios to the total number of images in the dataset. In Table 3 the summary of the VOC2012 dataset with a $\sigma = 0.05$ and $\beta = 1.4$ is given. The value of β was chosen empirically as the one which gave the least reconstruction error and the highest classification accuracy, and thus, can be said to have a balance between the number of outliers and inliers in the summary. It is evident from the table that our model trains in a way so that the percentage of images selected from each class is inline with the number of images of that class present in the complete dataset. For example, from table 3 we see that 2.58% of the dataset contains images of the class bicycle, and in the summary 2.41% of images are of the class bicycle, which is quite close to the initial composition. Similarly, for the person class, 40.02% images in the dataset and 41.01% images in the summary are present. Thus, we can say that the relevant information in an image corpus is maintained by our model by maintaining the composition of different classes in the summary. Thus, we can say that by minimizing the sparsity loss, regularize the composition of images of different classes in the summary and by minimizing the reconstruction loss, we ensure the most relevant images are included in this composition.

Outliers are images which are difficult to classify. There could be many reasons behind an image being an outlier, one being, its diversity. For example, consider an image containing a number of pets, say, a cat, a dog, a rabbit, and a horse. Now, no matter which one of the four labels is given to such an image, it will be an outlier, as the cross-entropy loss of such an image would be high because the probability that it belongs any of the remaining classes

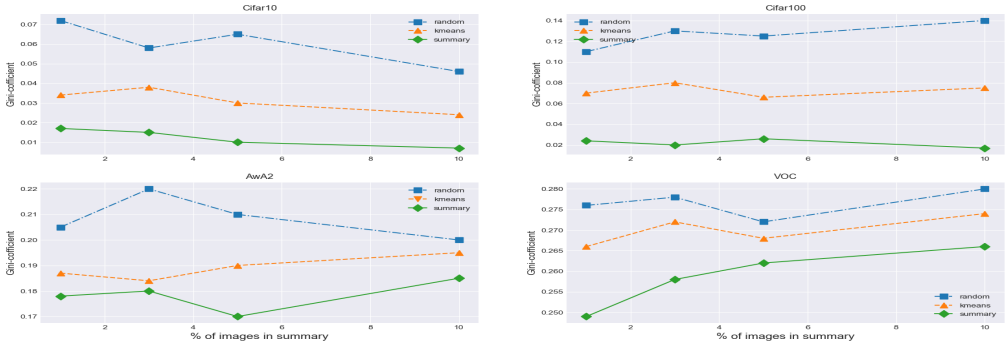


Figure 2: The gini coefficient for different number of images in summary of multiple datasets. With images selected at random from data set, selected using k-means clustering and images selected in summary from the model

Accuracy					
Method	Dataset	$\sigma = 1$	$\sigma = 0.1$	$\sigma = 0.3$	$\sigma = 0.5$
Ours	CIFAR10	88.1%	78%	83.8%	85.1%
	CIFAR100	66.2%	55.4%	60.6%	65.4%
	VOC2012	80.2%	76.3%	78.6%	79.1%
	AwA2	92.9%	89.5%	90.4%	91.9%
[1]	CIFAR10	82.1%	71%	78.8%	80.1%
	CIFAR100	60.2%	49.8%	56.6%	60.4%
	VOC2012	75.9%	70.1%	71.9%	76.3%
	AwA2	90.0%	82.1%	86.2%	88.9%
[3]	CIFAR10	84.1%	76%	80.7%	83.6%
	CIFAR100	65.1%	50.9%	56.8%	65.4%
	VOC2012	78.9%	72.3%	75.0%	77.2%
	AwA2	92.0%	84.1%	89.0%	90.9%

Table 2: Classification accuracy for different datasets

would be significant. Thus, we say that in order to ensure diversity, outliers must be present in the summary. In our paper, we show how task-specific loss is responsible for incorporating the outliers in the summary, and the affects of varying β on the number of outliers included in the summary.

For table 3, we calculate the number of outliers in each class of the VOC2012 dataset using a loss threshold of 0.1 i.e. all images having a cross-entropy loss greater than 0.1 were considered to be outliers. A truly diverse summary must have a composition similar to this table, mostly containing the outliers of different classes. the values in table 3 are calculated in the same way as table 3, except this time, β is taken to be a small value equal to 0.5. In the paper, we show that small values of β ensure the presence of a greater number of outliers in the summary. Here, we see that how small values of β lead to loss of relevant information from the summary by scaling up the affect of the task-specific loss, and leading the composition of the summary to be similar to the one only made up of outliers. For example, consider the class, dining-table. It's balanced composition, the one with the most appropriate number of outliers and inliers, in table 3, shows that 2.78% of the summary must comprise of the images of this class. But, when the β is lowered, its composition from table 3, shows that 4.27% of summary contains the images of this class. Thus, we see that by taking a significant affect of the task-specific with a lower value of β leads the composition of the summary to be similar to the composed of only outliers, as in table 3. Therefore, there is a trade-off between relevance and diversity in the summary.

VOC full dataset	Class	bicycle	motorbike	bird	airplane	horse
	No. of images	390	395	669	604	421
	Ratio	0.0258	0.0262	0.0443	0.0400	0.0279
	No. of outliers	143	130	26	24	140
	Ratio of outliers	0.0421	0.0382	0.0076	0.0071	0.041
	Class	car	tv-monitor	train	chair	cat
	No. of images	700	376	470	491	946
	Ratio	0.0463	0.0249	0.0311	0.0325	0.0626
	No. of outliers	156	122	51	281	55
	Ratio of outliers	0.0459	0.0359	0.0150	0.0826	0.016
	Class	cow	boat	dog	potted-plant	sheep
	No. of images	293	399	1069	234	303
	Ratio	0.0194	0.0264	0.0708	0.0155	0.0201
	No. of outliers	59	31	112	83	41
	Ratio of outliers	0.0174	0.0091	0.0329	0.0244	0.0121
	Class	bottle	sofa	bus	dining-table	person
	No. of images	342	337	360	261	6044
	Ratio	0.0226	0.0223	0.0238	0.0173	0.4002
	Number of outliers	149	215	39	169	1374
	Ratio of outliers	0.0438	0.0632	0.0115	0.0497	0.4041
Summary $\sigma = 0.05$ $\beta = 1.4$	Class	bicycle	motorbike	bird	airplane	horse
	No. of images	19	21	40	24	19
	Ratio	0.0241	0.0266	0.0506	0.0304	0.0241
	Class	car	tv-monitor	train	chair	cat
	No. of images	36	22	34	27	31
	Ratio	0.0456	0.0278	0.0430	0.0342	0.0392
	Class	cow	boat	dog	potted-plant	sheep
	No. of images	19	20	62	12	12
	Ratio	0.0241	0.0253	0.0785	0.0152	0.0152
	Class	bottle	sofa	bus	dining-table	person
	No. of images	14	16	16	22	324
	Ratio	0.0177	0.0203	0.0203	0.0278	0.4101
Summary $\sigma = 0.05$ $\beta = 0.5$	Class	bicycle	motorbike	bird	airplane	horse
	Number of images	13	14	28	26	19
	Ratio	0.0163	0.0176	0.0351	0.0326	0.0238
	Class	car	tv-monitor	train	chair	cat
	Number of images	28	37	17	56	23
	Ratio	0.0351	0.0464	0.0213	0.0703	0.0289
	Class	cow	boat	dog	potted-plant	sheep
	Number of images	20	8	26	15	11
	Ratio	0.0251	0.0100	0.0326	0.0188	0.0138
	Class	bottle	sofa	bus	dining-table	person
	Number of images	17	54	6	34	345
	Ratio	0.0213	0.0678	0.0075	0.0427	0.4329

Table 3: The number of images of each class for the VOC full dataset(15104 images) and their proportion in dataset. Also the Outliers (3400 images) on loss threshold of 0.1, and their ratio Similarly Number of images of a class in summary(790 images)at $\sigma = 0.05$ and $\beta = 1.4$ and their ratio to the total number of images in summary. Reconstruction Loss = 0.410. Classification Accuracy = 68.36. summary at $\sigma = 0.05$ and $\beta = 0.5$, and their ratio. Reconstruction Loss = 0.452. Classification Accuracy = 65.85

5. Visualization

5.1. t-SNE

We give the t-SNE visualization plots for different experiments conducted on the AWA2, VOC2012 and diversity 2016 data sets. For Diversity 2016, we generate image-embedded t-SNE plots for ground truth with top 20 ranked images from each class i.e. 1400 images in Figure (3)(a) and top 50 ranked images from each class i.e. 3500 images in Figure 3(c). We also give the corresponding set of 1400 and 3500 images generated by our model in Figure 3(b) and Figure 3(d). For these figures, we have highlighted some of the images which were selected by our model, and were also present in the ground truth summary. Further, t-SNE scatter plots as in figure 4(a), 4(b), 4(a) and 4(b) show how the summary generated from our models in more sparse and avoids clusters, as compared to the randomly generated summary.

5.2. Qualitative Ground Truth Comparison

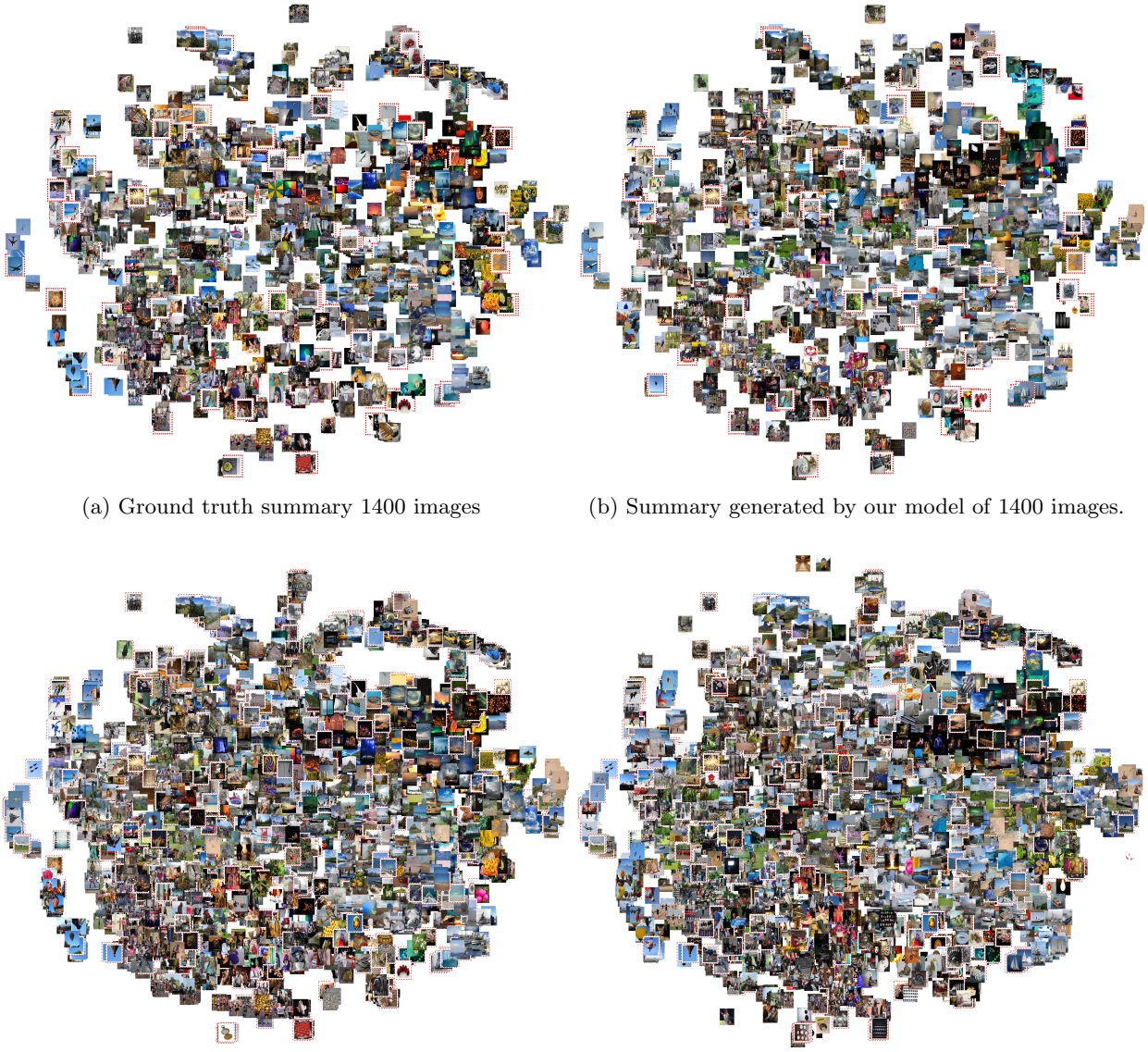
For qualitative assessment of the model proposed in this paper a comparison on diversity dataset was done. Where images were ranked according to their relevancy and diversity in a search result for a particular keyword which in turn was the class of those images. For every class we chose to keep the number of clusters 6. So the $K = 6 \times 70$ that is number of clusters in each class times the number of classes. In Figure 6 we show the top-6 results in the diversity dataset vs the 6 images selected from clustering by our model.There is intersection between images in the ground truth and the images selected by summary.

5.3. Tile Visualization

Due to lack of a standard dataset for task of generic summarization we thought of a novel method to qualitatively identify the efficacy of model. The qualitative visualization is done on a single image rather than a dataset where the image is broken into patches of size 32 pixels (tiny images). These tiny images are then fed forward to the model to generate feature vectors which are then clustered by the model. Depending upon the length of the summary, the number of tiny images are selected. In Figure 7 we use the famous lenna image commonly used in image processing research[32] and break it into tiny block images of size 32×32, then clustered for different σ. Within the Figure 7 it can be observed that the model tends to discard the background and keeps most relevant and diverse information blocks as the size of summary continues to decrease. The eyes and the blocks containing hair are not so similar in their feature representations as compared to two adjacent blocks in the background. The trade-off between relevance and diversity suggests that for some values of σ minimum redundancy and high diversity can be simultaneously achieved. Decreasing the σ further will cause the important parts of the dataset/image to be discarded and keeping the σ too high would mean keeping redundancy. The efficacy can be observed by comparison with randomly picked blocks which do not consider the relevance and diversity of a block. While a randomly generated summary will be most effused because of the random nature but it is likely to neglect the relevance of a block absolutely which can be seen in case of all 10%,30% and 50% of the blocks picked randomly.

6. Review of Task Specific Loss $\mathcal{L}_{task}$

Aim of task-specific loss is to generate task specific summary, in other words, task specific loss must be able to generate summary in cognizance of the intent by which summary is being produced. The intent is embedded into the objective by means of pre-trained network that is used and the pre-trained loss that is followed by that network. Although many diversity regularization losses like Detrimental Point Processes or Repelling Regularization Losses[1] try to achieve diversity. But most of diversity regularization works in unsupervised fashion to help select images with high visual diversity by selecting points that are sufficiently distant in feature space representation. Repelling regularization works in a way that it repels selection of images which are near to each other in the feature space. But there should be a priority in terms of selection of images from feature space where images of multiple labels cluster together in the dataset. Since diversity regularization techniques mentioned above are unsupervised it is less likely that they would take care of such scenarios where a outlier of a particular class rests within the cluster of inliers of another class. For the use case of task specific summarization it is necessary that such outliers do get picked in general along with ensuring similar high visual diversity. As task specific loss leverages labels it is able to distinguish between images



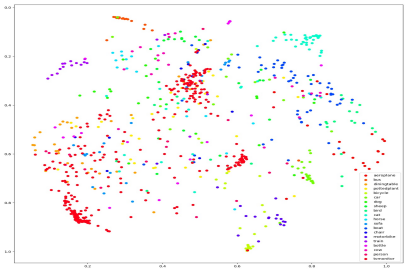
(a) Ground truth summary 1400 images

(b) Summary generated by our model of 1400 images.

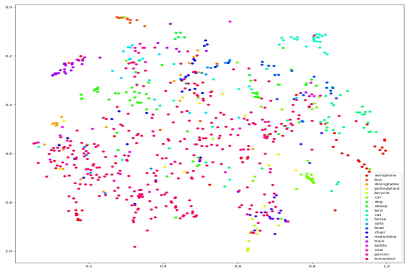
(c) Ground truth summary 3500 images

(d) 3500 images selected by our model

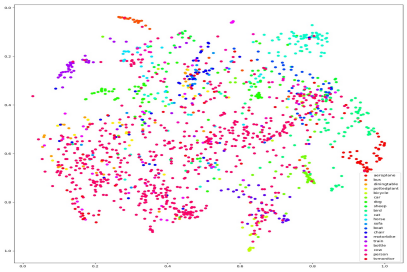
Figure 3: Karaparthy style t-SNE plot for Diversity2016 of our model in comparison to ground truth. The red boundaries highlight the images which are common to both summary and ground truth. Top-50 and Top-20 images selected from each of 70 classes to make ground truth of 3500 and 1400 images respectively. Please zoom in for better visualization



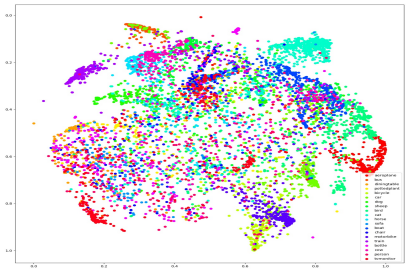
(a) VOC randomly selected 5% images



(b) VOC model selected 5% images

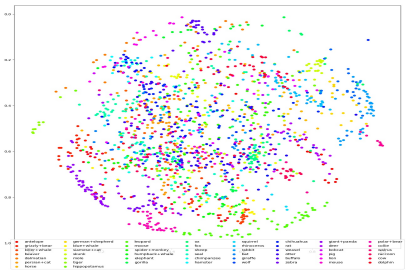


(c) VOC model selected 10% images

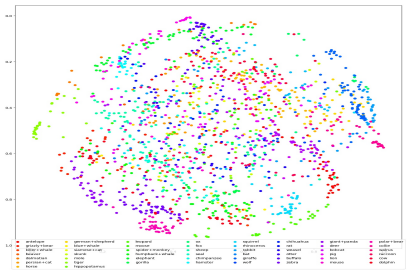


(d) VOC full dataset

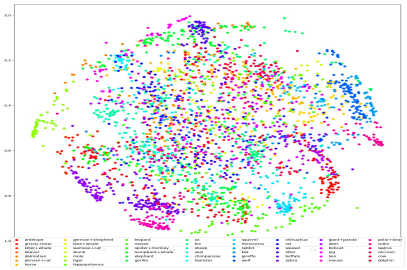
Figure 4: The above figure talks about the t-SNE plots for VOC data set Fig:(a) is a randomly generated summary for VOC dataset at 5%. Fig(b) and Fig(c) is the summary for VOC dataset, 5% and 10% of images selected from dataset by our model respectively. Fig:(d) is t-SNE for full voc dataset. Different colors represent different classes. Please zoom in for better visualization



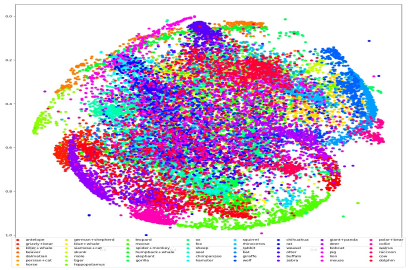
(a) AWA randomly selected 5% images



(b) AWA model selected 5% images



(c) AWA model selected 10% images



(d) AWA full dataset

Figure 5: The above figure talks about the t-SNE plots for AWA data set with Fig:(a) is a randomly generated summary for AWA dataset at 5%. Fig(b) and Fig(c) is the summary for dataset at 5% and 10% of images selected from dataset by our model. The Fig:(d) is t-SNE for full AwA dataset. Different colors represent different classes. Please zoom in for better visualization

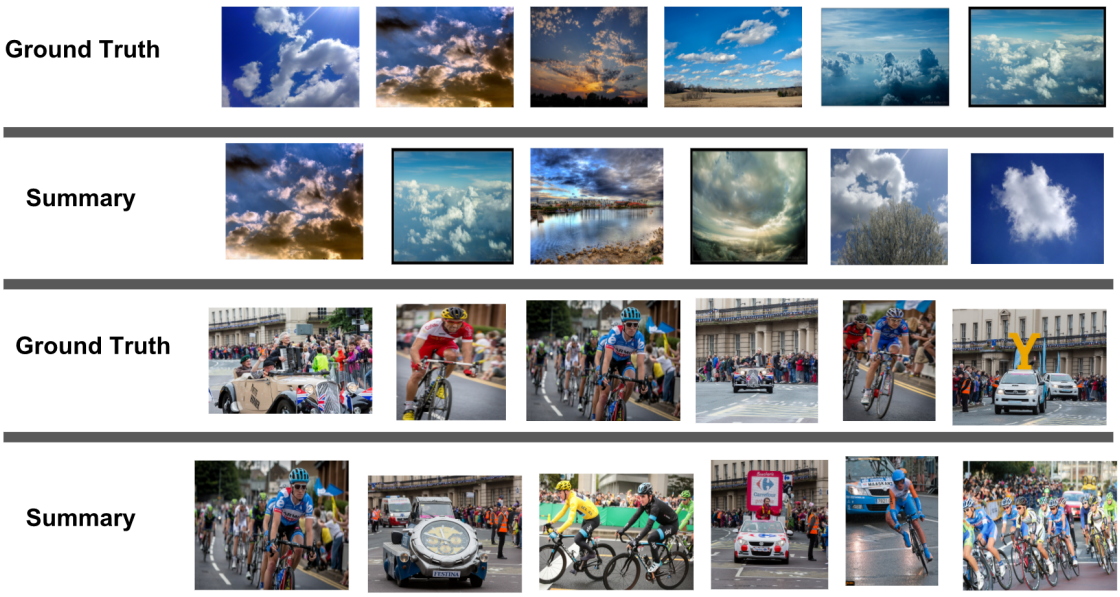


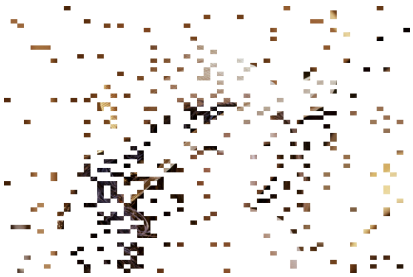
Figure 6: Comparison of the clouds and Tour-de France class images selected by our model ranked using ground truth vs top-6 images in ground truth for diversity 2016 dataset.

from class that are present within cluster of images of other class as outliers. It can be observed in the t-SNE visualization scatter plots for example in Figure 4(b) in the top right corner within selected images of a particular class i.e. cat we can find an outlier image of person being selected. Same can be observed in Figure 5(b) for example in bottom left corner where images from giant panda are selected outliers from other classes are also selected.

7. Conclusions and Future Works

In this work, we propose an unsupervised model to summarize a large collection of images. The classification results attained by training a deep network on summary only and on original dataset are close and show similar trend. Thus, model can also be used for a quick analysis of various models without needing to train on entire dataset. Moreover, image summarization can also be used to curate more precise data sets for given tasks. Summarization of data sets can be performed and accuracy achieved on summaries can be used to find out the small scale data sets for quicker training of models. In case the labels are not available, 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.

There is a need for a standard in the terms of dataset specific summarization both in the terms of general summarization and task specific summarization. There is no standard dataset that focuses on task of summarization. Evaluation in cases where ground truth reference is available is be done using F-score metrics. However, there is no standard and comprehensive metric when human annotated summary results are not available. Seeking solution to both problems will help build better supervised algorithms and also in testing efficacy of unsupervised models in a more rigorous manner. It was also observed that the idea of image collection summarization can help keep a strong foundation for image compression, saliency detection within images and attention models. The image collection summarization models can be fine tuned and used for the above mentioned tasks. For example similar idea to Figure 7 can perform saliency detection or attention model that it keeps only important parts in a image for small summary lengths i.e low values

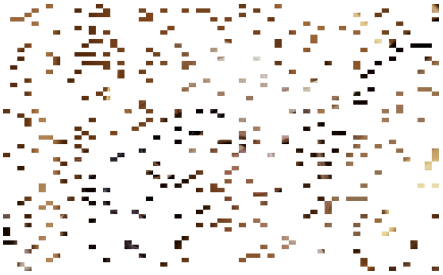


(a) 10% blocks selected by model

(b) 30% blocks selected by model

(c) 50% blocks selected by model

(d) 70% blocks selected by model



(e) 10% blocks selected randomly

(f) 30% of blocks selected randomly

(g) 50% of blocks selected randomly

(h) original image

Figure 7: Tile visuals on leena: Summary of the image for different values of σ and comparison with randomly generated summary. The figure shows a qualitative comparison of how proposed model is able to capture global semantics and is therefore able to preserve saliency of image.

of σ. Also if the discarded information can be estimated using the remaining summary. The reconstruction would act as the compressed image from lossy compression. There is also scope in exploring design of much better $\mathcal{L}_{sim}$ by exploring the manifold learning techniques to find the relationships and the overall latent high dimension plane in which all the feature embedding can be easily discriminated from each other.

[1] B. Mahasseni, M. Lam, S. Todorovic, Unsupervised video summarization with adversarial lstm networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2982–2991.

[2] K. Zhang, W.-L. Chao, F. Sha, K. Grauman, Video summarization with long short-term memory, in: Proceedings of the European conference on computer vision, 2016, pp. 766–782.

[3] K. Zhou, Y. Qiao, T. Xiang, Deep reinforcement learning for unsupervised video summarization with diversity-representativeness reward, in: Thirty-Second AAAI Conference on Artificial Intelligence, 2018.

[4] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: Advances in neural information processing systems, 2014, pp. 2672–2680.

[5] I. Simon, N. Snavely, S. M. Seitz, Scene summarization for online image collections, in: Proceedings of the IEEE International Conference on Computer Vision, 2007, pp. 1–8.

[6] J. E. Camargo, F. A. González, A multi-class kernel alignment method for image collection summarization, in: Iberoamerican Congress on Pattern Recognition, Springer, 2009, pp. 545–552.

[7] D. Stan, I. K. Sethi, eid: a system for exploration of image databases, Information processing & management 39 (3) (2003) 335–361.

[8] D. Deng, Content-based image collection summarization and comparison using self-organizing maps, Pattern Recognition 40 (2) (2007) 718–727.

[9] C. P. Chen, C.-Y. Zhang, Data-intensive applications, challenges, techniques and technologies: A survey on big data, Information Sciences 275 (2014) 314–347.

[10] C. Yang, J. Shen, J. Peng, J. Fan, Image collection summarization via dictionary learning for sparse representation, Pattern Recognition 46 (3) (2013) 948–961.

[11] A. Khosla, R. Hamid, C.-J. Lin, N. 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.

[12] S. Tschiatschek, R. K. Iyer, H. Wei, J. A. Bilmes, Learning mixtures of submodular functions for image collection summarization, in: Advances in neural information processing systems, 2014, pp. 1413–1421.

[13] D. Cai, X. He, Z. Li, W.-Y. Ma, J.-R. 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.

[14] B. Gao, T.-Y. Liu, T. Qin, X. Zheng, Q.-S. Cheng, W.-Y. 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.

[15] J.-Y. Chen, C. A. Bouman, J. C. Dalton, Hierarchical browsing and search of large image databases, IEEE transactions on Image Processing 9 (3) (2000) 442–455.

[16] P. Sinha, S. Mehrotra, R. Jain, Effective summarization of large collections of personal photos, in: Proceedings of the 20th international conference companion on World wide web, ACM, 2011, pp. 127–128.

[17] A. S. Virmani L, A. Subramanyam, Summary Generation for Image Corpus using Generative Adversarial Networks (2018).
URL `https://anurag14.github.io/images/Access_Template.pdf`

[18] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, CoRR abs/1512.00567. arXiv:1512.00567.
URL `http://arxiv.org/abs/1512.00567`

[19] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: 2009 IEEE conference on computer vision and pattern recognition, Ieee, 2009, pp. 248–255.

[20] D. J. Ketchen, C. L. Shook, The application of cluster analysis in strategic management research: an analysis and critique, Strategic management journal 17 (6) (1996) 441–458.

[21] R. L. Thorndike, Who belongs in the family?, Psychometrika 18 (4) (1953) 267–276.

[22] A. B. L. Larsen, S. K. Sønderby, H. Larochelle, O. Winther, Autoencoding beyond pixels using a learned similarity metric, arXiv preprint arXiv:1512.09300.

[23] A. Argyriou, T. Evgeniou, M. Pontil, Multi-task feature learning, in: Advances in neural information processing systems, 2007, pp. 41–48.

[24] S. Ruder, An overview of multi-task learning in deep neural networks, CoRR abs/1706.05098. arXiv:1706.05098.
URL `http://arxiv.org/abs/1706.05098`

[25] W. Chen, X. Chen, J. Zhang, K. Huang, A multi-task deep network for person re-identification, CoRR abs/1607.05369. arXiv:1607.05369.
URL `http://arxiv.org/abs/1607.05369`

[26] A. Krizhevsky, G. Hinton, Learning multiple layers of features from tiny images.

[27] Y. Xian, C. H. Lampert, B. Schiele, Z. Akata, Zero-shot learning - A comprehensive evaluation of the good, the bad and the ugly, CoRR abs/1707.00600. arXiv:1707.00600.
URL `http://arxiv.org/abs/1707.00600`

[28] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, A. Zisserman, The pascal visual object classes (voc) challenge, International Journal of Computer Vision 88 (2) (2010) 303–338.

[29] B. Ionescu, A. L. Gînscă, B. Boteanu, M. Lupu, A. Popescu, H. Müller, Div150multi: A social image retrieval result diversification dataset with multi-topic queries, in: Proceedings of the 7th International Conference on Multimedia Systems, 2016, pp. 46:1–46:6. doi:10.1145/2910017.2910620.
URL `http://doi.acm.org/10.1145/2910017.2910620`

[30] Y. Song, J. Vallmitjana, A. Stent, A. Jaimes, Tvsum: Summarizing web videos using titles, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 5179–5187.

[31] C. Gini, Variabilità e mutabilità, Reprinted in Memorie di metodologica statistica (Ed. Pizetti E, Salvemini, T).

Rome: Libreria Eredi Virgilio Veschi.
[32] Y. Fisher, Fractal image compression: theory and application, Springer Science & Business Media, 2012.