Radial Loss for Learning Fine-grained Video Similarity Metric

Abstract

In this paper, we introduce the Radial Loss that learns a fine-grained video similarity metric. In the presence of category and sub-category labels, the Radial Loss models the pairwise distance relationships between various types of video clips (set of video clip quadlets) to learn an order-preserving fine-grained video similarity metric. We propose an end-to-end quadlet-based Convolutional Neural Network (CNN) + Long Short-term Memory (LSTM) Unit based Recurrent Neural Network architecture that encodes variable length video clips into fixed length feature embeddings. These feature vectors are then passed to the Radial Loss to model the video similarities. We showcase two novel applications of learning a video similarity metric - (i) fine-grained video retrieval, (ii) fine-grained event detection, alongwith simultaneous shot boundary detection, and correspondingly show promising results against those of the baselines on two new video datasets. Additionally, we did cross domain study (on videos and images) to ensure the generalisability and the wide applicability of the proposed framework.

Introduction

The increased consumption of digital media in recent years has stemmed the need to develop algorithms, which automatically analyze video data for various applications such as video classification, retrieval, summarization, etc. Convolutional Neural Networks (CNNs) have set the benchmark results on different computer-vision tasks involving images (Krizhevsky, Sutskever, and Hinton 2012). In order to extend the utility of CNN's to perform different tasks on video data, multiple methods have been proposed in literature that include 3D convolutional networks (Hou, Chen, and Shah 2017), multi-resolution architecture utilizing local spatio-temporal information (Karpathy et al. 2014) and most notably, combining CNN's with Recurrent Neural Networks (RNNs) that utilize Long Short-Term Memory Units (LSTM) to capture video information both in the spatial and temporal domain (Donahue et al. 2015; Ng et al. 2015; Venugopalan et al. 2015). The CNNs learn the image level features, whereas the RNN with the LSTM architecture is shown to model the temporal dependency between these features and thereby enable the network to encode a variable length video to a fixed length feature, which is useful for video classification. However, such a training for the purpose of video classification results in feature embeddings that are only able to capture the category-level similarity (by learning categorical class specific features) and are generally not extendable to learn a fine-grained similarity (sub-class specific features) between a pair of videos. To learn a fine-grained image similarity, methods have been proposed in (Schroff, Kalenichenko, and Philbin 2015; Wang et al. 2014; Law, Thome, and Cord 2017; Weihua et al. 2017) to learn feature embedding through a *triplet-based* or *quadruplet-based* loss function. However, there is no work in the prior art that employs these methods for learning a fine-grained video similarity.

Learning a fine-grained video similarity allows for various applications. A cannonical example of such an application is fine-grained video retrieval(FgVR), where given a query video clip, a ranked list of similar video clips can be retrieved. An indirect application would be detecting fine-grained events within videos and simultaneously identifying shot boundaries. Specifically, for detecting critical events, which are fine-grained activities in a video that tend to increase user anxiety and anticipation. These activities could be a part of a whole video event and do not necessarily have defined shot boundaries resulting in very poor video frames classification accuracies for the conventional shot boundary detection (SBD) methods. Highlighted frames of Clip 2 shown in Fig. 2 illustrates an example of such a critical event. Both of these are novel applications that have not been attempted in the past.

In this paper, we introduce the *Radial Loss* that is able to learn a fine-grained video similarity metric. We argue for the improved feature embedding learned using the Radial loss over the triplet and quadruplet-based losses introduced in the prior art to learn image similarity. The primary distinction between the proposed loss with those in the prior art is the ability to learn an order-preserving fine-grained similarity by considering relative ordering between different types of video clip pairs. We propose a quadlet based CNN + LSTM network architecture that is able to learn the video embeddings through a set of video clip quadlets (A quadlet is consisting of query:q, positive:p, intermediate:i, negative:n video clips). The Radial loss captures the implicit relationships between video clip pairs such that $dist(q,p) < dist(q,i) < dist(q,n)$ & $dist(p,i) < dist(p,n)$ where $dist(a,b)$ denotes the distance between learned embeddings corresponding to video clips a and b. Using learned similarity metric, we

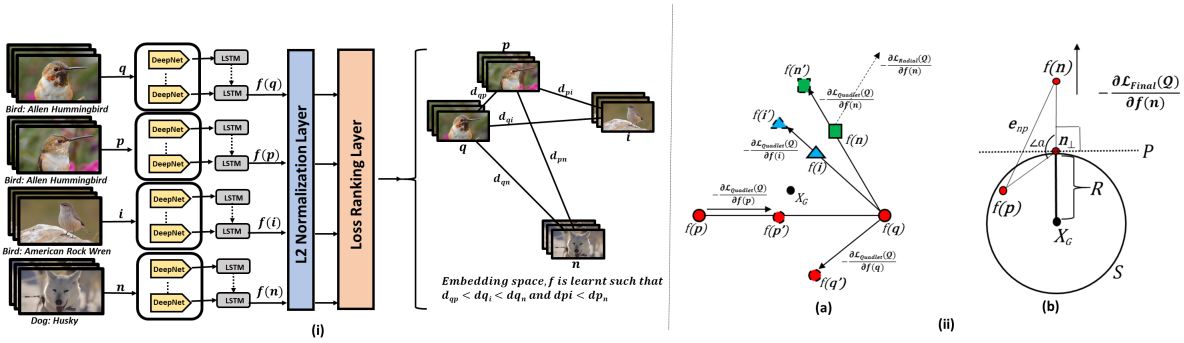

Figure 1: (i) Network architecture of proposed framework. (ii)(a) Illustration of quadlet loss and its gradient directions (relative to position of f(q)) on a synthetic example for which $\mathcal{L}_1, \mathcal{L}_2 > 0$ and $\mathcal{L}_3 = 0$. (ii)(b) Illustration of Theorem 3.1 ($\angle a > 90°$)

calculate L2 norm difference of the different pairs to identify critical clips and shot boundaries, and use the same to retrieve videos in the task of order ranking of FgVR. To the best of our knowledge, this is the first attempt to perform Critical Clip Detection (CCD) and FgVR using similarity learning.

The main contributions of our work are as follows:

1. We introduce a generic supervised deep similarity learning based framework using quadlet based CNN + LSTM architecture capable of (a) performing simultaneous shot boundary and critical clip detection, and (b) learning order-preserving fine-grained video similarity metric.
2. We introduce the novel *Radial loss*, which learns video similarity metric to perform the aforementioned tasks. We also propose the *quadlet loss*, which is an intuitive extension of triplet loss for the quadlets and show their comparison.
3. We create (a) soccer highlight videos dataset with clip and critical clip annotations for SBD and CCD; (b) Dog-Birds dataset to evaluate our proposed methodology for FgVR.
4. We showcase robustness of proposed framework with its superior performance against state-of-the-art baselines for the task of order preserving fine-grained image retrieval.

Related Work

Traditional methods to learn a video similarity employ hand-crafted features (Jiang, Ngo, and Yang 2007; Patel and Meshram 2012). Recently, deep learning based solutions that utilize a Siamese network have been proposed to learn a similarity metric between time-series data (Mueller and Thyagarajan 2016; Pei, Tax, and van der Maaten 2016). Also, methods have been proposed in the literature to learn feature embeddings through triplet-based or quadruplet-based loss function (Schroff, Kalenichenko, and Philbin 2015; Law, Thome, and Cord 2017; Chen et al. 2017) for the purpose of learning fine-grained similarity in images, that are known to outperform the Siamese network architecture. Additionally, to improve upon results achieved by triplet-based loss (Sohn 2016) proposes the multi-class N-pair loss that jointly learns from multiple negative samples at every batch as opposed to a single negative sample within triplet-loss and (Oh Song et al. 2016) proposes a triplet based Lifted Structured Similarity Softmax loss, which uses sophisticated triplet sampling strategy to outperform standard triplet-based loss. A variation of the triplet-loss was proposed by (Wang et al. 2017) which takes into account the angle relationship between triplet samples to show improved performance. Although these methods are shown to improve the performance of triplet loss to learn fine-grained image similarity through different sampling strategies, similar attempts have not been made in prior art to learn a fine-grained video similarity. Literature consists of many fine-grained action detection and segmentation papers (Lea et al. 2016; Lea, Vidal, and Hager 2016; René and Hager 2017) but none of them focus on FgVR and thus, only limited multi-class datasets are available. Though there are video datasets with fine-grained annotations such as 50 Salads(Stein and McKenna 2013), VB100 (Ge et al. 2016), JIGSAWS (Gao et al. 2014), each of them has videos belonging to one class (Cooking, Birds etc.) with further fine sub-class distinction. Thus, we exploit VB100 and YouTube to create our own multi-class dataset for FgVR. For CCD, we obtain annotations and present results on a dataset of soccer highlights, but the proposed framework can easily be adapted to other notions of criticality. In the next section we discuss the Radial loss formulation along with details of the quadlet network.

Proposed Methodology

We propose a quadlet-based network to learn video embedding function for detecting shot boundaries, identifying critical clips and for fine-grained video retrieval task (see Fig. 1(i)). The network takes a quadlet $\mathcal{Q}$ as an input consisting of query sample (q), positive sample (p), intermediate sample (i) and negative sample (n). In general, each sample consists of a set of consecutive video frames that are randomly picked from the annotated videos during each training iteration. Altogether, quadlets are selected from the training dataset and fed into four identical deep neural networks with shared architecture and parameters (see Fig1). The term DeepNet in this paper refers to the existing trained deep neural network AlexNet (Krizhevsky, Sutskever, and Hinton 2012). Each DeepNet is followed by an LSTM cell at the top to learn long-term dependencies for each sample in $\mathcal{Q}$. Lastly, the output of these networks is fed into L2-normalization layer. A loss ranking layer then evaluates the proposed loss over the

Figure 2: Temporally contiguous clips segmented by hard cuts. Clip-2 captures the same scene left off at the end of Clip-1 but from a different angle. Highlighted frames (yellow border) of Clip-2 are the critical frames where the ball hit by the player enters the net. Clip-3 is a close up celebration shot.

set of training quadlets. During learning, it tries to preserve the relative ordering across the samples in selected quadlet.

The goal is to find the feature embeddings such that (a) p and n lie far apart in the embedding space, and (b) i lies near to p with some margin which is smaller to the margin between p and n. Whenever there is a violation in the ordering of quadlets, the loss is calculated and the gradients are computed over the loss and back-propagated to the lower layers. The lower layers adjust their parameters to minimize the proposed loss to learn an order-preserving fine-grained ranking. For the task of fine-grained image retrieval, we have a similar quadlet-based network but without any LSTM cells. Thus, selected quadlets consisting of *(q,p,i,n)* images are fed into four identical DeepNets followed by L2-normalization layer.

Formally, the goal is to learn embedding function $f(.)$ such that following order preserving fine-grained relations (see Eqns. 1 and 2) exist for a given quadlet, $\mathcal{Q} = (q_k, p_k, i_k, n_k)$:

$$dist(f(q), f(p)) < dist(f(q), f(i)) < dist(f(q), f(n)) \quad (1)$$

$$\text{and } dist(f(p), f(i)) < dist(f(p), f(n)) \quad (2)$$

$$\text{where, } dist(f(a), f(b)) = ||f(a) - f(b)||_2^2$$

To ensure properties given in Eqns. 1 and 2, we first state the Quadlet Loss, and then extend it to a more robust Radial Loss to mitigate its caveats.

For the two tasks of FgVR and CCD we demonstrate two different strategies for quadlet selection (q, p, i, n video clips) for training. For the task of video retrieval, given the class and sub-class labels for videos, selection is done in such a way that q and p share the same fine-grained category while q and i share the same category/class at a coarser level but different sub-categories at a fine-grained level. Lastly, q and n belong to different classes. For the task of identifying fine-grained events within videos, we demonstrate our selection strategy with an example. A soccer highlight video (shown in Fig.2) is comprised of critical (Clip-2) and non-critical clips (Clip-1,3) segregated by hard-cuts, where criticality of a clip is dictated by the events that take place in the clip. For example, player kicking the ball towards the goal constitute a critical event shown in highlighted frames (yellow) of Clip-2. We find a stark similarity between the highlighted frames and rest of the frames within Clip-2 as compared to those of Clip-1 and 3. Thus, q and p are selected from non-critical frames (from non-highlighted frames of Clip-2 in Fig.2) and i is selected from the frames annotated as “critical" within the same clip (highlighted frames of Clip-2 in Fig.2). n is selected from the next consecutive clip (Clip-3 in Fig.2).

Quadlet Loss

The triplet loss($\mathcal{L}_1$) proposed in (Wang et al. 2014), models $dist(f(q_k), f(p_k)) < dist(f(q_k), f(n_k))$ by pushing the negative point $f(n_k)$ away from anchor point $f(q_k)$ by a distance margin $g_1 > 0$ as compared to positive point $f(p_k)$. To model other relations in Eq. 1, 2, a natural extension would be to consider hinge function to model the losses in the following manner:

$$\mathcal{L}_1(q_k, p_k, n_k) = max\{0, g_1 + ||f(q_k) - f(p_k)||_2^2 - ||f(q_k) - f(n_k)||_2^2\} \quad (3)$$

$$\mathcal{L}_2(q_k, p_k, i_k) = max\{0, g_2 + ||f(q_k) - f(p_k)||_2^2 - ||f(q_k) - f(i_k)||_2^2\} \quad (4)$$

$$\mathcal{L}_3(p_k, i_k, n_k) = max\{0, g_3 + ||f(p_k) - f(i_k)||_2^2 - ||f(p_k) - f(n_k)||_2^2\} \quad (5)$$

$$\mathcal{L}_{Quadlet}(\mathcal{Q}) = \mathcal{L}_1(q_k, p_k, n_k) + \mathcal{L}_2(q_k, p_k, i_k) + \mathcal{L}_3(p_k, i_k, n_k) \quad (6)$$

where, g_1, g_2 & g_3 denote the gap parameters that regularizes the gaps between the image pairs $\{(q_k, p_k)$ & $(q_k, n_k)\}$, $\{(q_k, p_k)$ & $(q_k, i_k)\}$, $\{(p_k, i_k)$ & $(p_k, n_k)\}$ respectively such that $g_2 \leq g_3 \leq g_1$. The intuition being that the gap between (n_k, i_k), should be more than that between (p_k, i_k) and hence $g_3 \geq g_2$. Similarly, margin between (p_k, n_k) should be more than (i_k, n_k) and hence $g_1 \geq g_3$. To optimize Eq. 6, we calculate its gradient with respect to each sample of the quadlet respectively as:

$$\frac{\partial \mathcal{L}(\mathcal{Q})}{\partial f(.)} = \begin{cases} [2(f(q) - f(i))]_{\mathcal{L}_2>0} + [2(f(i) - f(p))]_{\mathcal{L}_3>0}, & w.r.t \;\; f(i) \\ [2(f(q) - f(n))]_{\mathcal{L}_1>0} + [2(f(p) - f(n))]_{\mathcal{L}_3>0}, & w.r.t \;\; f(n) \\ [2(f(n) - f(p))]_{\mathcal{L}_1>0} + [2(f(i) - f(p))]_{\mathcal{L}_2>0}, & w.r.t \;\; f(q) \\ [2(f(p) - f(q))]_{\mathcal{L}_1>0} + [2(f(p) - f(q))]_{\mathcal{L}_2>0} + [2(f(n) - f(i))]_{\mathcal{L}_3>0}, & w.r.t \;\; f(p) \end{cases} \quad (7)$$

The back propagation estimates the embedding function $f(.)$ in such a way that the relationships of the quadlet loss

in Eqns. 1 and 2 are preserved. But, it can be observed that the derived gradient fails to consider the interaction with other points. Consider, the negative point, $f(n)$ for example. Gradient either moves it away from $f(p)$ ($\mathcal{L}_3 > 0$) or $f(q)$ ($\mathcal{L}_1 > 0$) or from local sub-class distribution of $f(p)$ & $f(q)$ i.e from its center $\frac{f(p)+f(q)}{2}$ (when $\mathcal{L}_{1,3} > 0$). Its gradient may not be optimal unless without the guarantee of moving away from overall class distribution which query $f(q)$, positive $f(p)$ and intermediate $f(i)$ samples belong to.

Radial Loss

The constraints imposed by Quadlet loss drive $f(n)$ away from $f(p)$ & $f(q)$ but can bring it dangerously close to $f(i)$ as shown in Fig 1.(ii)(a). This result contradicts with our goal to keeping discrete class samples far apart in the embedding space. Intuitively, similar to the work in (Wang et al. 2017) we can derive constraints to push $f(n)$ away from local cluster defined by $\mathcal{T} = (f(q), f(p), f(i))$ and model the relation between $f(n)$ and $\mathcal{T}$. A natural approximation to this distribution is a 3D sphere, $\mathcal{S}$ centered at the centroid X_G of triangle $\triangle f(q)f(p)f(i)$, and radius, R large enough to engulf $\mathcal{T}$. Mathematically, the Radial loss would consist of minimizing the following:

$$\mathcal{L}_{radial}(\mathcal{Q}) = \Big[(cR)^2 - ||f(n) - X_G||_2^2\Big]_+ \tag{8}$$
$$X_G = (f(q) + f(p) + f(i))/3,$$
$$R = \max\{||f(q) - X_G||_2, ||f(p) - X_G||_2, ||f(i) - X_G||_2\}$$

i.e. $f(n)$ should be at least at a distance of cR from the center of the sphere. Such value of R always keeps at least one of the points in $\mathcal{T}$ on the surface and rest of them inside the sphere. c is the multiplier that decides the margin between $f(n)$ and X_G and its value should be such that $f(n)$ is sufficiently far enough from all points in $\mathcal{T}$. We use the following theorem to decide the margin, cR by getting a lower bound on c.

Theorem 0.1. *If $\mathcal{L}_{radial}(\mathcal{Q}) = 0$ using $c \geq 3$, then both Eq(1) and Eq(2) hold, consequently achieving sufficient separation of $f(n)$ from $\mathcal{T}$.*

Proof: Assuming that we have minimized Radial loss such that $\mathcal{L}_{radial}(\mathcal{Q}) = 0$ which implies:-

$$\begin{aligned} & dist(f(n), X_G) \geq cR \geq 3R \\ \Rightarrow\ & dist(f(n), n_\perp) + dist(n_\perp, X_G) \geq 3R \\ \Rightarrow\ & dist(f(n), n_\perp) \geq 2R \\ & \big(\because dist(n_\perp, X_G) = R, \text{See Fig } 1.(ii)(b)\big) \end{aligned} \tag{9}$$

where, $n_\perp$ is the orthogonal projection of $f(n)$ on the sphere $\mathcal{S}$ such that for any vector $\vec{r}$ lying on the tangent plane, P we have $(\vec{r} - \vec{n_\perp}) \cdot (\vec{f(n)} - \vec{n_\perp}) = 0$. Since, any two points on or inside a sphere can have a maximum separation of $2R$, we get

$$\{dist(f(q), f(p)),\ \ dist(f(p), f(i)),\ \ dist(f(q), f(i))\} \leq 2R \tag{10}$$

Now, consider point $f(p)$ and construct $\triangle f(p)n_\perp f(n)$ as shown in Fig 1.(ii)(b). Since, $f(p)$ lies inside or on the sphere, $\angle a > 90^\circ$ is an obtuse angle which makes side opposite to it, e_{np} the longest.

$$\text{Thus, } e_{np} = dist(f(p), f(n)) > dist(f(n), n_\perp) \tag{11}$$
$$\text{(Using (9), (10), (11), we obtain)}$$
$$\Rightarrow dist(f(p), f(n)) > dist(f(p), f(i))$$

Similarly, for point $f(q)$, we can prove $dist(f(q), f(p)) < dist(f(q), f(n))$ & $dist(f(q), f(i)) < dist(f(q), f(n))$ in Eq1. Lastly, to consider $dist(f(q), f(p)) < dist(f(q), f(i))$ we will simply add a triplet loss to $\mathcal{L}_{radial}(\mathcal{Q})$ to obtain the following final loss:

$$\mathcal{L}_{Radial}(\mathcal{Q}) = \mathcal{L}_{radial}(\mathcal{Q}) + \mathcal{L}_{triplet}(\mathcal{T}) \tag{12}$$
$$\mathcal{L}_{triplet}(\mathcal{T}) = max\{0, m + ||f(q) - f(p)||_2^2 - ||f(q) - f(i)||_2^2\}$$

To better understand the effect of optimizing this loss, we compute its gradient with respect to $f(n)$[1] and identify that the gradient pushes $f(n)$ radially outwards away from the center of local cluster, X_G defined by $\mathcal{T}$ by considering interaction with $f(q), f(p), f(i)$ (see Fig 1(ii) (b)):

$$-\partial\mathcal{L}_{Radial}(\mathcal{Q})/\partial f(n) = 2(f(n) - X_G) \tag{13}$$

Having learned the embeddings, for (a) **Simultaneous shot boundary and critical clip detection**, (i) positive and negative clips (different clips) lie far in space (margin m), which helps to detect the shot boundaries in a video; and (ii) intermediate clips (critical) lie near to positive clip with some margin m' ($< m$), which helps to detect critical event within a detected clip (if it exists). For (b) **Fine-grained video retrieval**, consider the following scenario where we have a two class database *[C_1: Dog; C_2: Bird]* with 2 subclasses each *[Dog_{B1}, Dog_{B2}, $Bird_{B1}$, $Bird_{B2}$]*. Given a query from the subclass Dog_{B1}, the relative margins between samples help in realizing following ranking in retrieved sequence in terms of relevance: $Dog_{B1} > Dog_{B2} > Bird_{B1/B2}$. Similarly, learned margins help in realizing correct ranking order in the task of fine-grained image retrieval.

Datasets

To evaluate our approach, we use the following datasets[2]: (a) Sports dataset for the simultaneous clip and critical clip detection, (b) Dog-Birds(DB) dataset for FgVR and (c) CnD, FtW, FcP datasets for FgIR.

Sports Dataset: It consists of 25 annotated Youtube club soccer highlights that are conceptually similar but visually different in terms of lighting conditions during the game (daytime, night, bright sunlight and rainy), camera shots (close-up, medium or long shot). The conceptual similarity is dictated by the universal rules of the game "football". The collected videos also share some visual similarities such as green color of the football field. Each highlight video is approximately 2 minutes in length. A total of 400 clips are annotated and extracted at 20 fps. Out of them, 180 clips contain critical events which are observed to appear at the end of every clip.

Soccer highlights are annotated based on the duration of shots/clips where a shot contains all the temporally contiguous frames as shown in Fig.2. A shot boundary is decided

[1]Rest of the gradients are included in Supplementary Material.

[2]Dataset Link: *https://gofile.io/?c=8aphlD*

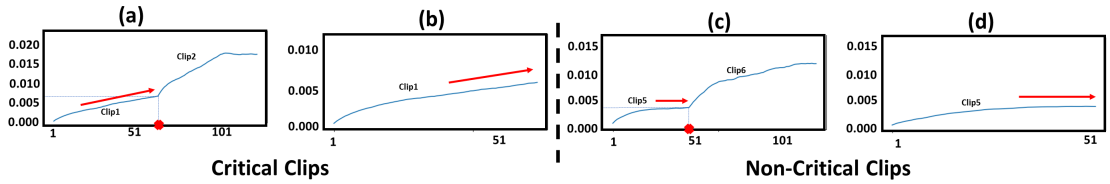

Figure 3: Graphs (a) and (c) are first used to determine hard-cuts ($(N_o - 1)$ shown in red dots after which frames from next clip start appearing in the window). Then, graphs (b) and (d) show the trend ('increasing' for critical or 'flat' for non-critical) in the identified Clips (Clip-1,5). The Window ID is plotted on the X-axis and L2 difference on Y-axis.

based on the presence/absence of hard cut. For **CCD**, based on the definition of criticality we label those frames from a clip as "critical" that correspond to the event when the ball is hit in the direction of goal (highlighted frames in Fig.2). Due to the inherent subjectivity in the task, it is infeasible to obtain clear ground truth labels for critical frames. Thus, the annotation was done by 10 users in two steps. Firstly, it was identified whether clip contained the critical event or not. Secondly, frames corresponding to the critical event were annotated by the majority of users who labeled the clip as critical. In case of inter-annotator disagreement, only the frames with maximum agreement amongst users were considered as critical. We opted for highlights of soccer videos to make the task of SBD and CCD challenging owing to the camera movement to keep the ball in view, jump cuts, multi-view camera shots of events like goals scored/missed, fouls etc.

Dog-Birds Dataset: Due to unavailability of annotated videos for fine-grained video retrieval, we constructed our own train-test dataset. It consists of 307 videos each 15-20 secs long from two classes:- (1) Dogs which have 9 sub-classes (135 videos) and (2) Birds which have 12 sub-classes (172 videos). Each sub-class has on average 10 training videos and 5 test videos. A Frame rate of 20fps is used. The Birds dataset is borrowed from 100-classes VB100 Video Bird Dataset(Ge et al. 2016) (first 12 sub-classes in alphabetical order) while Dogs dataset has been built by the authors of the paper. The videos pose several challenges such as variations in scale, pose, background, camera movement etc.

FgIR Dataset: We have worked on following image datasets for Fine-grained image retrieval- **(i) Cats and Dogs, CnD** (Parkhi et al. 2012): It consists of 7384 images with class labels = *[Dogs, Cats]* and subclass labels = *25 dog and 12 cat breed names*, **(ii) Footwear, FtW** (Yu and Grauman 2014): This dataset consists of 50025 images with class labels = [Shoes, Sandals, Slipper, Boots] and subclass labels = *21 functional footwear types* and lastly, **(iii) Face Pose, FcP** (fac) consisting of 1890 images. Here, the class labels are assigned from *[Front, Left, Right]* whereas, the 90 subjects form the subclass labels for fine-grained categorization.

Results and Evaluation

This section presents the results and evaluation of proposed model for different tasks.

Evaluation Metrics

Shot Boundary and Critical Clip Detection (SBD & CCD) : We have used Precision, Recall and F1 to measure the quality of both hard-cut and CCD; in which for SBD, TPs, are correctly marked cuts, FPs are the number of undetected cuts and FNs are number of falsely detected cuts. While for CCD, precision is the probability that an assumed clip by the model is actually a critical clip while recall is the fraction of existing critical clips detected by the model.

Fine-grained Video Retrieval (FgVR): Following metrics are used to capture order-preserving fine-grained and coarse-grained level similarity:

Order-preserving fine-grained video/clip similarity: (a) *Quadlet Precision* (QP) is defined as the percentage of quadlets being correctly ranked. A given quadlet, $Q = (q, p, i, n)$, is correctly ranked if $p > i > n$, i.e. p is more similar to q than i and n, and i is more similar to q as compared to n, but less similar than p. (b) *Normalized Discounted cumulative gain* (NDCG) measures the usefulness, or gain, of a clip based on its rank in the retrieved list. The gain is accumulated from the top of the result list to the bottom, with the gain of each retrieved sample discounted at lower ranks. Order-preserving fine-grained image ranking is a three-class problem, where we have samples from positive, intermediate and negative class.

Coarse grained video/clip similarity: (a) *Triplet Precision* (TP) relaxes the hard constraint of order preserving across positive and intermediate samples. It is defined as the percentage of triplets being correctly ranked. A triple $T = (q, p, n)$ is correctly ranked, if the ranking order is $p > n$. (b) *Mean Average Precision* (MAP) is the mean of average precision, across all the test/query samples. For each query sample in the test set, we obtain the ranked retrieval list using our learned similarity metric. It considers the positive and intermediate samples belong to the same class.

We have used similar metrics to capture coarse-grained and fine-grained image similarity but with corresponding notions of images.

Experimental Settings

Training: For the simultaneous learning of shot boundaries and critical clip detection, we have used the AlexNet Model. For fully connected layer, we have used: 512 filters, Xavier as weight filler, learning rate multiplier $= 10$ and decay rate multiplier $= 1$. For quadlet loss we have used the following margin parameters: $g_1 = 0.7$, $g_2 = 0.3$, and $g_3 = 0.5$, determined experimentally giving best separation between the sample types. For Radial Loss, we set $m = 0.3$(empirically determined) for $\mathcal{L}_{triplet}$ and $c = 3$ leading to overall convergence of the Radial Loss. It is observed that first triplet loss

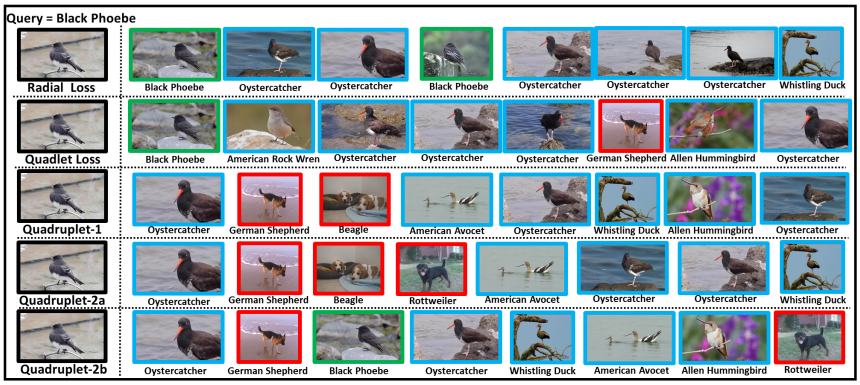

Figure 4: Visual illustration of retrieved ranking order for proposed and baseline methods (Quadruplet 1, 2a, 2b). Red box represents the irrelevant samples (*negatives*), green box represents the relevant samples (*positives*), and blue box represents the intermediately relevant samples (*intermediates*).

converges fixing $cR = 3||f(i) - X_G||_2$ for every iteration, then $\mathcal{L}_{radial}$ eventually converges. For CCD and FgVR, each sample in $\mathcal{Q}$ consists of 40 frames and 100 frames respectively. In CCD, quadlets are generated for critical clips as explained in "Introduction"; while for non-critical clips(Clip-1 in Fig2) triplets are generated (q, p from Clip-1 and n from consecutive Clip-2) and corresponding Triplet Loss, $\mathcal{L}_{triplet}(f(q), f(p), f(n))$ with margin 0.7 is minimized.

Testing: For simultaneous detection of shot/clip boundaries and critical clips, we first identify hard-cuts in a video. For this, we perform a sliding window operation (stride = 1, window size=40 frames) on the video to generate the encoding of the frames within the window. We then analyze the plot of $||f_1 - f_n||_2^{n>1}$; where, f_1 is the feature encoding of first 40 frames of a clip, n is the window ID and f_n is the encoding of n^{th} window (shown in Fig.3). At $n = N_o$, sudden change in the slope is observed due to appearance of first frame from the consecutive clip (see Fig.3(a),(c)). Thus, a hard-cut exists between last two frames of window N_o. Continuing in a similar fashion for the remaining video, we reset the f_1 to first window of the clip following a hard-cut. The trend in the plot of identified clip ($||f_1 - f_n||_2^{1<n<N_o}$) is then analyzed to determine if the clip is critical (as shown in Fig.3(b),(d)). It is observed in the dataset that a critical event usually appears at the end of the clip. Thus, an increasing trend determines that the clip is critical ($\because dist(f_q, f_i) > dist(f_q, f_p)$), where f_q, f_p and f_i are feature embeddings of windows constituting query, positive and intermediate samples respectively. For non-critical clips with no intermediate samples, positive samples should lie at similar distances from query ($\because dist(f_q, f_{p_1}) \sim dist(f_q, f_{p_2})$) obtained in Fig.3(c), (d).

For testing fine-grained video/image retrieval framework, we have used four different metrics. To calculate QP, we have created 15K quadlets from the test split for the two datasets. To calculate TP, we have created the triplets using the same quadlets generated for QP. Each quadlet is expanded into two triplets. For e.g. given quadlet Q = (q; p; i; n), the corresponding triplets are T_1 = (q; p; n) and T_2 = (q; i; n). Thus, we have 30K (15K*2) triplets for evaluation. Similarly for ranking task (NDCG and MAP), we have used the test split as the query set and train split as the retrieval set.

Baselines

Shot Boundary Detection: Traditionally, methods have relied on hand-crafted features (G. and S. 2014; Lu and Shi 2013; Zhang and Wang 2012), spatial information and intensity histograms (Fang, Jiang, and Feng 2006; Lu and Shi 2013; Zhang and Wang 2012) to detect shot boundaries. Deep learning based solutions proposed recently, (Hassanien et al. 2017b; Gygli 2017) try to improve the detection accuracy. We have used following SBD baselines for comparison:

- In ***PySceneDetect*** (pys), a threshold of 30 is used for the content-aware detector(unsupervised).
- ***ShotDetect*** (sho) is also a unsupervised content-based detection method. A threshold of 30 is used.
- ***SVD-SBD*** (Lu and Shi 2013) uses intensity histogram to calculate frame similarity in HSV color space.
- ***DeepSBD*** (Hassanien et al. 2017a) is a supervised technique based on spatio-temporal CNNs. Segments of length 16 with an overlap of 8 are fetched to a deep 3D-CNN and are then assigned either of two labels: 1) sharp transition, 2) no transition since our dataset has no gradual transitions.

Fine-grained Video Retrieval:

- ***3D-CNN*** (Tran et al. 2015): The last fc layer of the 3D-CNN network was used for feature extraction. The network considering all sub-classes as discrete was trained for 21-class video classification.
- ***Quadruplet-1 (IJCV-2017)*** (Law, Thome, and Cord 2017): Using loss function and quadruplet sampling strategy proposed by (Law, Thome, and Cord 2017), we train the loss ranking layer of the proposed network Fig.1.
- ***Quadruplet-2 (CVPR-2017)*** (Weihua et al. 2017): We used the loss functions proposed by (Weihua et al. 2017) and generate samples in two different ways - where negative samples come from (a) both negative & intermediate categories, and (b) only from the negative class.
- ***Triplet*** (Wang et al. 2014): We have fine tuned the last layer of AlexNet model using the concept of triplet learning.

Fine-grained Image Retrieval: We have used Triplet, Quadruplet-1 and Quadruplet-2 as baselines.

	Prec	Recall	F1(%)
PySceneDetect	47.95	80.77	60.17
ShotDetect	49.10	73.0	48.70
SVD-SBD	38.30	62.0	47.35
DeepSBD	51.60	85.13	64.25
Triplet	80.50	85.13	82.75
Quadlet Loss(Ours)	85.00	87.93	86.44
Radial Loss(Ours)	**86.67**	**88.89**	**87.77**

Table 1: Model performances in Shot Boundary Detection

	QP(%)	NDCG	TP(%)	MAP
3D-CNN	25.89	68.65	57.36	45.45
Triplet	46.62	63.91	73.9	45.86
Quadruplet-1	51.20	66.07	75.61	46.14
Quadruplet-2a	50.90	65.99	75.80	45.99
Quadruplet-2b	50.85	64.03	75.64	44.19
Quadlet Loss(Ours)	52.42	69.34	75.03	46.88
Radial Loss(Ours)	**60.70**	**69.80**	**78.12**	**48.08**

Table 2: Model performances in Fine-grained video retrieval

	CnD				FcP				FtW			
	QP%	NDCG	TP%	MAP	QP%	NDCG	TP%	MAP	QP%	NDCG	TP%	MAP
Triplet	69.19	0.95	86.69	0.80	70.45	0.89	85.32	0.54	42.75	0.88	75.91	0.65
Quadruplet-1	78.14	0.95	90.99	0.84	79.77	0.90	88.78	0.64	37.12	0.86	71.92	0.59
Quadruplet-2a	76.90	0.94	89.91	0.85	77.65	0.90	88.52	0.61	45.88	0.87	76.13	0.66
Quadruplet-2b	51.83	0.92	87.38	0.82	78.42	0.90	87.06	0.60	38.64	0.84	74.77	0.61
Quadlet(Ours)	81.82	0.97	93.92	0.91	83.74	0.91	91.74	0.66	53.93	0.89	80.56	0.67
Radial(Ours)	**84.39**	**0.98**	**95.86**	**0.93**	**87.21**	**0.92**	**92.64**	**0.70**	**55.8**	**0.90**	**81.91**	**0.68**

Table 3: Quantitative metrics for order-preserving image ranking task (QP, NDCG) and coarse-grained ranking task (TP, MAP).

Results and Discussion

Shot Boundary Detection: Triplet, Quadruplet (baselines) and Quadlet network (ours) based approaches behave similarly for shot boundary detection, since there is no notion of intermediates in hard-cut detection (query and positive samples from same clip and negative samples from different clip). But the Quadlet results slightly differ because the Quadlet network was trained for simultaneous clip and critical clip detection which was not possible for triplet and quadruplet baselines. Our proposed model addresses the mentioned challenges in the sports dataset and has the highest precision and recall of all the models (Table 1). Low precision in baselines is observed due to the excess number of false positives that are detected by them mostly in medium shots due to significant scene changes in the same clip. Although the Quadlet Loss and Radial Loss perform similarly in SBD, we can ascertain the superiority of the Radial Loss in CCD and FgVR.
Critical Clip Detection: In the absence of existing baselines for the novel task of CCD, we compare the results of our two proposed losses. For CCD, Quadlet Loss achieves Precision=51.02%, Recall=**73.53**%, F1=60.24% while Radial Loss achieves Precision=**55.81**%, Recall=72.72%, F1=**63.15**%. In Fig.3, via graphs we demonstrate the procedure used to compute results for SBD and CCD.
Fine-grained Video/Image Retrieval: Using the NDCG and QP, in Table 2 and 3 we showcase superior performance of Radial Loss in order-preserving fine-grained ranking task. The baselines perform poorly because they either learn fine-grained categorization(3D-CNN) or consider pairwise losses to separate positive and negative samples from each other. Hence, their optimization function does not preserve the order across the fine-grained categories as shown in Fig.4. They only capture inter and intra-class variability utilizing the relative ordering between positive and negative pairs.

We tried to extend the work of (Wang et al. 2014) as a baseline by simultaneously minimizing three angular losses between triplets $T_1 = (q_k, p_k, n_k)$, $T_2 = (q_k, p_k, i_k)$, $T_3 = (p_k, i_k, n_k)$ to constrain the angles at the i_k of T_2 and n_k of T_1, T_3. This required defining relations between upper bounds α_1, α_2 and α_3 in T_1, T_2, T_3 respectively. Though, we can define $\alpha_1 < \alpha_2$ since we want n_k to be at a larger distance from both q_k, and p_k than i_k; we cannot define such relations of α_1, α_2 w.r.t α_3. Thus, Angular loss formulation fails to preserve the relative ordering between (q, p, i, n). Experimentally, we validated this hypothesis and found that angular loss did not converge for learning similarity between quadlets.

The Quadlet and Radial loss try to ensure that the relevancy is maintained even in lower rank-orders. However, gradients of the Quadlet loss might bring n close to i samples which can be seen in Fig.4 where an irrelevant sample although at a lower rank is still higher than some intermediate relevant samples. The proposed Radial Loss ensures higher separation between intermediate and negative samples, thus achieving significant qualitative and quantitative improvement. The second metric NDCG helps to capture the ranking relevance for multi-class problem by assigning a high relevance to the same class, low relevance to less relevant classes, and a zero relevance to the irrelevant classes. From Table 2, 3, proposed frameworks are performing better than the baselines for NDCG values. We use TP and MAP to show the effectiveness of the proposed framework in the coarse-grained retrieval task where the aim is to maintain the order between the q, p and n. From Table 2 and 3, we can see that for both TP and MAP ranking, our frameworks perform better than the existing baselines.

Conclusions and Future work

In this paper, we have proposed an end-to-end quadlet based CNN-LSTM network which uses Radial Loss to learn fine-grained similarity metric. The approach is validated on three tasks: Fine-grained Video/Image Retrieval and Critical Clip Detection. The proposed methodology outperforms baseline models in clip segmentation, achieves high performance in novel tasks CCD, FgVR. Furthermore, results of FgIR show robustness of the proposed Radial Loss.

References

Chen, W.; Chen, X.; Zhang, J.; and Huang, K. 2017. Beyond triplet loss: a deep quadruplet network for person re-identification. In *CVPR*.

Donahue, J.; Anne Hendricks, L.; Guadarrama, S.; Rohrbach, M.; Venugopalan, S.; Saenko, K.; and Darrell, T. 2015. Long-term recurrent convolutional networks for visual recognition and description. In *CVPR*.

Database for face detection and pose estimation.

Fang, H.; Jiang, J.; and Feng, Y. 2006. A fuzzy logic approach for detection of video shot boundaries. *CVPR*.

G., L. P. G., and S., D. 2014. Walsh hadamard transform kernel-based feature vector for shot boundary detection. *IEEE Transactions on Image Processing*.

Gao, Y.; Vedula, S. S.; Reiley, C. E.; Ahmidi, N.; Varadarajan, B.; Lin, H. C.; Tao, L.; Zappella, L.; Béjar, B.; Yuh, D. D.; et al. 2014. Jhu-isi gesture and skill assessment working set (jigsaws): A surgical activity dataset for human motion modeling. In *M2CAI*.

Ge, Z.; McCool, C.; Sanderson, C.; Wang, P.; Liu, L.; Reid, I.; and Corke, P. 2016. Exploiting temporal information for dcnn-based fine-grained object classification. In *Digital Image Computing: Techniques and Applications (DICTA)*.

Gygli, M. 2017. Ridiculously fast shot boundary detection with fully convolutional neural networks. *CoRR*.

Hassanien, A.; Elgharib, M.; Selim, A.; Hefeeda, M.; and Matusik, W. 2017a. Large-scale, fast and accurate shot boundary detection through spatio-temporal convolutional neural networks. *arXiv*.

Hassanien, A.; Elgharib, M. A.; Selim, A.; Hefeeda, M.; and Matusik, W. 2017b. Large-scale, fast and accurate shot boundary detection through spatio-temporal convolutional neural networks. *CoRR*.

Hou, R.; Chen, C.; and Shah, M. 2017. An end-to-end 3d convolutional neural network for action detection and segmentation in videos. *arXiv*.

Jiang, Y.-G.; Ngo, C.-W.; and Yang, J. 2007. Towards optimal bag-of-features for object categorization and semantic video retrieval. In *Image and video retrieval*.

Karpathy, A.; Toderici, G.; Shetty, S.; Leung, T.; Sukthankar, R.; and Fei-Fei, L. 2014. Large-scale video classification with convolutional neural networks. In *CVPR*.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. In *NIPS*.

Law, M. T.; Thome, N.; and Cord, M. 2017. Learning a distance metric from relative comparisons between quadruplets of images. *International Journal of Computer Vision*.

Lea, C.; Reiter, A.; Vidal, R.; and Hager, G. D. 2016. Segmental spatiotemporal cnns for fine-grained action segmentation. In *ECCV*.

Lea, C.; Vidal, R.; and Hager, G. D. 2016. Learning convolutional action primitives for fine-grained action recognition. In *ICRA*.

Lu, Z. M., and Shi, Y. 2013. Fast video shot boundary detection based on svd and pattern matching. *IEEE Transactions on Image Processing*.

Mueller, J., and Thyagarajan, A. 2016. Siamese recurrent architectures for learning sentence similarity. In *AAAI*.

Ng, J. Y.-H.; Hausknecht, M.; Vijayanarasimhan, S.; Vinyals, O.; Monga, R.; and Toderici, G. 2015. Beyond short snippets: Deep networks for video classification. In *CVPR*.

Oh Song, H.; Xiang, Y.; Jegelka, S.; and Savarese, S. 2016. Deep metric learning via lifted structured feature embedding. In *CVPR*.

Parkhi, O.; Vedaldi, A.; Zisserman, A.; and Jawahar, C. 2012. Cats and dogs. In *CVPR*.

Patel, B., and Meshram, B. 2012. Content based video retrieval systems. *arXiv*.

Pei, W.; Tax, D. M.; and van der Maaten, L. 2016. Modeling time series similarity with siamese recurrent networks. *arXiv*.

Pyscenedetect. `https://pyscenedetect.readthedocs.io/en/latest/`.

René, C. L. M. D. F., and Hager, V. A. R. G. D. 2017. Temporal convolutional networks for action segmentation and detection. *arXiv*.

Schroff, F.; Kalenichenko, D.; and Philbin, J. 2015. Facenet: A unified embedding for face recognition and clustering. In *CVPR*.

Shot detect. `https://github.com/johmathe/Shotdetect`.

Sohn, K. 2016. Improved deep metric learning with multi-class n-pair loss objective. In *NIPS*.

Stein, S., and McKenna, S. J. 2013. Combining embedded accelerometers with computer vision for recognizing food preparation activities. In *Ubicomp*.

Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; and Paluri, M. 2015. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*.

Venugopalan, S.; Rohrbach, M.; Donahue, J.; Mooney, R.; Darrell, T.; and Saenko, K. 2015. Sequence to sequence-video to text. In *ICCV*.

Wang, J.; Leung, T.; Rosenberg, C.; Wang, J.; Philbin, J.; Chen, B.; Wu, Y.; et al. 2014. Learning fine-grained image similarity with deep ranking. *CVPR*.

Wang, J.; Zhou, F.; Wen, S.; Liu, X.; and Lin, Y. 2017. Deep metric learning with angular loss. In *ICCV*.

Weihua, C.; Xiaotang, C.; Jianguo, Z.; and Kaiqi, H. 2017. Beyond triplet loss: a deep quadruplet network for person re-identification. In *CVPR*.

Yu, A., and Grauman, K. 2014. Fine-grained visual comparisons with local learning. In *CVPR*.

Zhang, C., and Wang, W. 2012. A robust and efficient shot boundary detection approach based on fisher criterion. In *ICM*.

Radial Loss for Learning Fine-grained Video Similarity Metric - Supplementary Material

Anonymous Author(s)
Affiliation
Address
email

1 Quadlet Loss

Using the definition of quadlet loss defined in the paper:

$$\mathcal{L}_{Quadlet}(\mathcal{Q}) = \mathcal{L}_1(q_k, p_k, n_k) + \mathcal{L}_2(q_k, p_k, i_k) + \mathcal{L}_3(p_k, i_k, n_k) \tag{1}$$

We can formulate our constrained optimization problem as:

$$\begin{aligned} &\text{minimize} \sum_k \xi_k + \sum_k \zeta_k + \sum_k \varrho_k + \lambda||W||_2^2 \\ &\text{subjected to:} \quad \mathcal{L}_1(q_k, p_k, n_k) \leq \xi_k, \quad \mathcal{L}_2(q_k, p_k, i_k) \leq \zeta_k, \quad \mathcal{L}_3(p_k, i_k, n_k) \leq \varrho_k \end{aligned} \tag{2}$$

where, W is the embedding learned by the CNN, and $f(.)$ denotes the function which projects a sample to the embedded space. λ is the regularization parameter that improves the generalization capability of the ranking algorithm. We use $\lambda = .001$ in this paper. The above constrained optimization problem can be converted into the following unconstrained optimization problem:

$$\text{minimize} \;\; \lambda||W||_2^2 + \sum_k \mathcal{L}_1(q_k, p_k, n_k) + \sum_k \mathcal{L}_2(q_k, p_k, i_k) + \sum_k \mathcal{L}_3(p_k, i_k, n_k) \tag{3}$$

2 Radial Loss Gradients

In this section, we compute gradients of the Radial Loss with respect to $f(q)$, $f(p)$, $f(i)$ and $f(n)$.

$$\partial\mathcal{L}_{Radial}(\mathcal{Q}) = \partial\mathcal{L}_{radial}(\mathcal{Q}) + \partial\mathcal{L}_{triplet}(\mathcal{T}) \tag{4}$$

Using $c = 3$ and assuming $R = ||f(i) - X_G||_2$ we obtain the following before the obtained $\mathcal{Q}$ is normalized,

Final Radial Loss Components

$$\mathcal{L}_{radial}(\mathcal{Q}) = 9R^2 - ||f(n) - X_G||_2^2 \tag{5}$$

$$\begin{aligned} \mathcal{L}_{radial}(\mathcal{Q}) = \frac{35}{9}f(i)^2 - f(n)^2 + \frac{8}{9}f(q)^2 + \frac{8}{9}f(p)^2 - \frac{38}{9}f(i)\Big(f(p) + f(q)\Big) \\ + \frac{2}{3}f(n)\Big(f(i) + f(p) + f(q)\Big) + \frac{16}{9}f(q)f(p) \end{aligned} \tag{6}$$

$$\mathcal{L}_{triplet}(\mathcal{T}) = ||f(q) - f(p)||_2^2 - ||f(q) - f(i)||_2^2 + m \tag{7}$$

$$\mathcal{L}_{triplet}(\mathcal{T}) = f(p)^2 - f(i)^2 - 2f(p)f(q) + 2f(i)f(q) + m \tag{8}$$

Submitted to 32nd Conference on Neural Information Processing Systems (NIPS 2018). Do not distribute.

Gradients of the final Radial Loss

$$\frac{\partial \mathcal{L}_{Radial}(\mathcal{Q})}{\partial f(q)} = \Big[\frac{16}{9}f(q) - \frac{38}{9}f(i) + \frac{2}{3}f(n) + \frac{16}{9}f(p)\Big]_{\mathcal{L}_{radial}>0} + \Big[2(f(i) - f(p))\Big]_{\mathcal{L}_{triplet}>0} \quad (9)$$

$$\frac{\partial \mathcal{L}_{Radial}(\mathcal{Q})}{\partial f(p)} = \Big[\frac{16}{9}f(q) - \frac{38}{9}f(i) + \frac{2}{3}f(n) + \frac{16}{9}f(p)\Big]_{\mathcal{L}_{radial}>0} + \Big[2(f(p) - f(q))\Big]_{\mathcal{L}_{triplet}>0} \quad (10)$$

$$\frac{\partial \mathcal{L}_{Radial}(\mathcal{Q})}{\partial f(i)} = \Big[-\frac{38}{9}f(q) + \frac{70}{9}f(i) + \frac{2}{3}f(n) - \frac{38}{9}f(p)\Big]_{\mathcal{L}_{radial}>0} + \Big[2(f(q) - f(i))\Big]_{\mathcal{L}_{triplet}>0} \quad (11)$$

$$\frac{\partial \mathcal{L}_{Radial}(\mathcal{Q})}{\partial f(n)} = \Big[2(X_G - f(n))\Big]_{\mathcal{L}_{radial}>0} \quad (12)$$

After normalization of the obtained feature vectors and ignoring constant terms, we obtain the following:

$$\mathcal{L}_{radial}(\mathcal{Q}) = -\frac{38}{9}f(i)\Big(f(p) + f(q)\Big) + \frac{2}{3}f(n)\Big(f(i) + f(p) + f(q)\Big) + \frac{16}{9}f(q)f(p) \quad (13)$$

$$\mathcal{L}_{triplet}(\mathcal{T}) = -2f(p)f(q) + 2f(i)f(q) + m \quad (14)$$

$$\frac{\partial \mathcal{L}_{Radial}(\mathcal{Q})}{\partial f(q)} = \Big[-\frac{38}{9}f(i) + \frac{2}{3}f(n) + \frac{16}{9}f(p)\Big]_{\mathcal{L}_{radial}>0} + \Big[2(f(i) - f(p))\Big]_{\mathcal{L}_{triplet}>0} \quad (15)$$

$$\frac{\partial \mathcal{L}_{Radial}(\mathcal{Q})}{\partial f(p)} = \Big[-\frac{38}{9}f(i) + \frac{2}{3}f(n) + \frac{16}{9}f(p)\Big]_{\mathcal{L}_{radial}>0} + \Big[-2f(q)\Big]_{\mathcal{L}_{triplet}>0} \quad (16)$$

$$\frac{\partial \mathcal{L}_{Radial}(\mathcal{Q})}{\partial f(i)} = \Big[-\frac{38}{9}f(q) + \frac{2}{3}f(n) - \frac{38}{9}f(p)\Big]_{\mathcal{L}_{radial}>0} + \Big[f(q)\Big]_{\mathcal{L}_{triplet}>0} \quad (17)$$

$$\frac{\partial \mathcal{L}_{Radial}(\mathcal{Q})}{\partial f(n)} = \Big[2X_G\Big]_{\mathcal{L}_{radial}>0} \quad (18)$$

Similarly, we can compute the loss and gradients when $R = ||f(q) - X_G||_2$ and $R = ||f(p) - X_G||_2$.