RADIAL LOSS FOR LEARNING FINE-GRAINED VIDEO SIMILARITY METRIC

Abhinav Jain[⋆‡] *Prerna Agarwal*[⋆‡] *Shashank Mujumdar*[⋆‡] *Nitin Gupta*[⋆‡] *Sameep Mehta*[⋆]
Chiranjoy Chattopadhyay[†]

[⋆] IBM Research Laboratory, India
[†] Indian Institute of Technology, Jodhpur

ABSTRACT

In this paper, we propose the Radial Loss which utilizes category and sub-category labels to learn an order-preserving fine-grained video similarity metric. We propose an end-to-end quadlet-based Convolutional Neural Network (CNN) combined with Long Short-term Memory (LSTM) Unit to model video similarities by learning the pairwise distance relationships between samples in a quadlet generated using the category and sub-category labels. We showcase two novel applications of learning a video similarity metric - (i) fine-grained video retrieval, (ii) fine-grained event detection, along with simultaneous shot boundary detection, and correspondingly show promising results against those of the baselines on two new fine-grained video datasets.

Index Terms— Fine-grained Video Similarity, CNN-LSTM, Radial Loss, Critical Event Detection

1. 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. Learning a fine-grained video similarity opens up space for following applications: (i) fine-grained video retrieval (FgVR), where given a query video clip, a ranked list of similar video clips can be retrieved, (ii) detecting critical events within videos and simultaneously identifying shot boundaries. Critical events are fine-grained activities in a video that tend to increase user anxiety and anticipation. These activities could be part of a whole video event and do not necessarily have defined shot boundaries resulting in poor video frames classification accuracy for the conventional shot boundary detection (SBD) methods. Highlighted frames of Clip 2 shown in Fig. 1 illustrate an example of a critical event in soccer highlight video.

In this paper, we propose the *Radial Loss*, capable of learning a fine-grained video similarity metric. The Radial loss captures implicit relationships between samples of a quadlet (consisting of query:q, positive:p, intermediate:i, negative:n samples) such that $D(f(q), f(p)) < D(f(q), f(i)) < D(f(q), f(n))$ and $D(f(p), f(i)) < D(f(p), f(n))$ where $D(f(a), f(b)) = ||f(a) - f(b)||_2^2$

Fig. 1: Illustration of Critical Event. Highlighted frames (yellow border) of Clip-2 are critical frames where the ball hit by the player enters the net.

We propose a deep quadlet based CNN + LSTM network architecture that is able to learn video embedding through a set of video clip quadlets. Using the overall framework, we learn a fine-grained video similarity metric 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 performing Critical Clip Detection (CCD) and FgVR using video similarity learning. We also argue for the improved feature embedding learned using the Radial loss over the triplet and quadruplet-based losses introduced in the prior art and showcase its superior results in the aforementioned applications.

2. RELATED WORK

Traditional methods to learn a video similarity employ hand-crafted features [1, 2]. Recently, deep learning based solutions have been proposed in the literature to learn feature embeddings through triplet-based or quadruplet-based loss function [3, 4, 5] for the purpose of learning fine-grained similarity in images. A variation of the triplet-loss was proposed in [6] which takes into account the angle relationship between triplet samples to show improved performance. Although these methods improve the performance of triplet loss to learn fine-

[‡] Authors contributed equally to the work

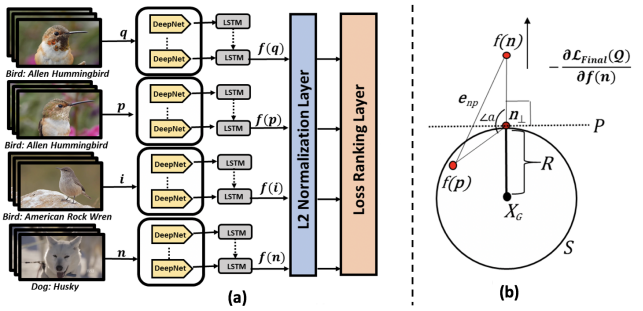

Fig. 2: (a) Network architecture of proposed framework. (b) Illustration of Theorem 3.1 ($\angle a > 90°$)

grained image similarity through different sampling strategies, there is no work in prior art that employs such methods for learning fine-grained similarity in videos.

Literature consists of fine-grained action detection and segmentation work, [7, 8, 9] none of them focussing on FgVR; thus, no multi-class datasets with fine-grained annotations such as 50 Salads[10], VB100 [11], JIGSAWS [12]. We exploit VB100 and YouTube to create our own multi-class dataset for FgVR. For CCD, we obtain annotations and present results on soccer highlights dataset, but the proposed framework can easily be adapted to other notions of criticality.

3. PROPOSED METHODOLOGY

We propose a quadlet-based framework (see Fig. 2(a)) that takes a quadlet $\mathcal{Q} = \{q, p, i, n\}$ as input where each sample consists of a set of consecutive video frames that are picked from the annotated video dataset during each training iteration. The term DeepNet in this paper refers to the existing trained deep neural network AlexNet [13]. 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 a L2-normalization layer. A loss ranking layer then evaluates the proposed loss over the set of training quadlets. During learning, it tries to preserve the relative ordering across the samples in selected quadlet.

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 refer to Fig. 1. In the soccer highlight video a player kicking the ball towards the goal constitutes a critical event as shown in highlighted frames (yellow) of Clip-2. It is separated from Clip-1,3 by hard-cuts which are non-critical. We can see 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.1) and i is selected from the frames annotated as "critical" within the same clip (highlighted frames of Clip-2 in Fig.1). n is selected from the next consecutive clip (Clip-3 in Fig.1).

3.1. Radial Loss

Our aim is to formulate a loss such that given a quadlet $\mathcal{Q} = \{q, p, i, n\}$ - (i) p and n lie far in space (margin m) (ii) i lie near to p with some margin $m'(< m)$ (see Fig. 2(b)). 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:

$$\mathcal{L}_{radial}(\mathcal{Q}) = \Big[(cR)^2 - ||f(n) - X_G||_2^2\Big]_+ \tag{1}$$

where $R = \max\{||f(q) - X_G||_2, ||f(p) - X_G||_2, ||f(i) - X_G||_2\}$ and $X_G = (f(q)+f(p)+f(i))/3$. Thus, $f(n)$ 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 3.1. *If $\mathcal{L}_{radial}(\mathcal{Q}) = 0$ using $c \geq 3$, achieves 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:-

$$\begin{aligned} & D(f(n), X_G) \geq cR \geq 3R \\ \Rightarrow\ & D(f(n), n_\perp) + D(n_\perp, X_G) \geq 3R \\ \Rightarrow\ & D(f(n), n_\perp) \geq 2R \end{aligned} \tag{2}$$

$\because dist(n_\perp, X_G) = R$, See Fig. 2($b$) 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

$$max\{D(f(q), f(p)), D(f(p), f(i)), D(f(q), f(i))\} \leq 2R \tag{3}$$

Now, consider point $f(p)$ and construct $\triangle f(p)n_\perp f(n)$ as shown in Fig. 2(b). Since, $f(p)$ lies inside or on the sphere, $\angle a > 90°$ is an obtuse angle which makes side opposite to it, e_{np} the longest.

$$\text{Thus, } e_{np} = D(f(p), f(n)) > D(f(n), n_\perp) \tag{4}$$

$$\Rightarrow D(f(p), f(n)) > D(f(p), f(i)) \quad \{\text{From (2), (3), (4)}\}$$

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

$$\mathcal{L}_{Radial}(\mathcal{Q}) = \mathcal{L}_{radial}(\mathcal{Q}) + \mathcal{L}_{triplet}(\mathcal{T}) \quad (5)$$
$$\mathcal{L}_{triplet}(\mathcal{T}) = [m + ||f(q) - f(p)||_2^2 - ||f(q) - f(i)||_2^2]^+$$

To better understand the effect of optimizing this loss, we compute its gradient with respect to $f(n)$ 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. 2(b)):

$$-\partial\mathcal{L}_{Radial}(\mathcal{Q})/\partial f(n) = 2(f(n) - X_G) \quad (6)$$

4. DATASETS

To evaluate our approach, we use the following datasets[1]: (a) Sports dataset for the simultaneous clip and critical clip detection, (b) Dog-Birds(DB) dataset for FgVR

Sports Dataset: It consists of 2 min long 25 annotated Youtube club soccer highlights. A total of 400 clips (180 critical clips) are annotated and extracted at 20 fps. For **CCD**, frames corresponding to the event when the ball is hit towards the goal (highlighted frames in Fig. 1) constitute the 'critical event'. Due to involved inherent subjectivity in the task, annotation was done by 10 users. In case of inter-user disagreement, only the frames with maximum agreement amongst users were retained as critical.

Dog-Birds Dataset: It consists of 307 videos each 15-20 secs long from two classes:- (1) Dogs (9 sub-classes, 135 videos) and (2) Birds (12 sub-classes, 172 videos). Each sub-class has on average 10 training videos and 5 test videos. 20 fps is used.

5. RESULTS AND EVALUATION

5.1. Evaluation Metrics

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 in terms of similarity if $p > i > n$. (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.

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.

Shot Boundary and Critical Clip Detection (SBD & CCD): We have used Precision, Recall and F1 to measure the quality of both SBD and CCD; where, 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.

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

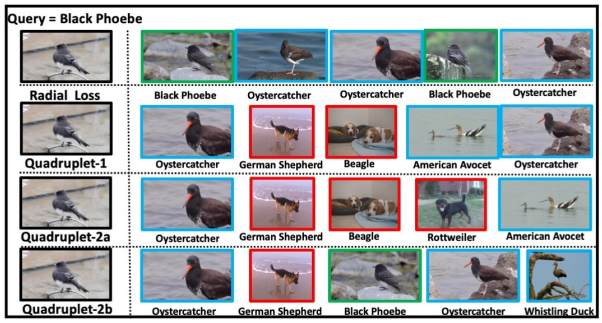

Fig. 3: Visual illustration of retrieved ranking order for proposed and baseline methods. Red box represents the irrelevant samples (*negatives*), green box represents the relevant samples (*positives*), and blue box represents the intermediately relevant samples (*intermediates*).

5.2. Experimental Settings

Training: For the AlexNet model, we have used: 512 filters for fully connected layer, Xavier as weight filler, learning rate multiplier = 10 and decay rate multiplier = 1. We have empirically set margin parameters: $g_1 = 0.7$, $g_2 = 0.3$, $g_3 = 0.5$, $m = 0.3$ and $c = 3$ giving best separation between the samples. For CCD and FgVR, each sample in $\mathcal{Q}$ consists of 40 frames and 100 frames respectively. In CCD, quadlets are generated (see Sec. 3) when q is sampled from 'critical clip' and triplets are generated (q, p from same clip and n from next consecutive clip) when q is sampled from 'non-critical clip'. In latter, $\mathcal{L}_{Radial} = \mathcal{L}_1(q, p, n)$ 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. 4). At $n = N_o$, sudden change in the slope is observed due to appearance of first frame from the consecutive clip (see Fig. 4(a),(c)). Thus, a hard-cut

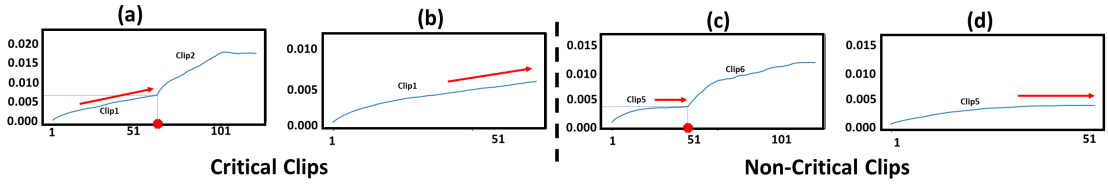

Fig. 4: Graphs (a) and (c) are first used to determine hard-cuts ($(N_o - 1)$ shown in red dots after which frames from next clip start appearing in the window). Then, graphs (b) and (d) show the trend ('increasing' for critical or 'flat' for non-critical) in the identified Clips (Clip-1,5). The Window ID is plotted on the X-axis and L2 difference on Y-axis.

	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
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
Radial Loss (Ours)	**60.70**	**69.80**	**78.12**	**48.08**

Table 2: Model performances in Fine-grained video retrieval

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.4(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. 4(c), (d).

5.3. Baselines

We utilize the following baselines for comparison along with specifications mentioned in parentheses for the two tasks.
Shot Boundary Detection: (i) *PySceneDetect* [14] (threshold = 30), (ii) *ShotDetect* [15] (threshold = 30), (iii) *SVD-SBD* [16], (iv) *DeepSBD* [17] (segment length=16, overlap=8).
Fine-grained Video Retrieval: (i) *3D-CNN* [18], (ii) *Triplet* [19], (iii) *Quadruplet-1* [4] (use loss function and quadruplet sampling strategy), (iv) *Quadruplet-2(a,b)* [20] (use loss function and two sampling strategies - where negative samples come from (a) both negative & intermediate categories, and (b) only from the negative class).

5.4. Discussion

Shot Boundary Detection: Proposed framework 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 mostly in shots with significant scene changes in the same clip. However, proposed framework significantly outperforms the baselines for SBD due to better separation in the embedding space.
Critical Clip Detection: In absence of existing baselines for the novel task of CCD, we report the results of proposed framework using Radial Loss - Precision=**55.81**%, Recall=72.72%, F1=**63.15**%. In Fig. 4, via graphs we demonstrate the procedure used to compute results for SBD and CCD.
Fine-grained Video Retrieval: Using NDCG and QP, in Table 2 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 without factoring in the intermediate samples. Hence, their optimization function does not preserve the order across the fine-grained categories as shown in Fig. 3. Radial loss tries to ensure that the relevancy is maintained even in lower rank-orders achieving higher separation between intermediate and negative samples achieving high NDCG values shown in Table 2. 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, we can see that for both TP and MAP ranking, our framework performs better than the existing baselines.

6. 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 video similarity metric. The approach is validated on two novel tasks: Fine-grained Video Retrieval and Critical Clip Detection. The proposed methodology outperforms baseline models in clip segmentation and achieves high performance in novel tasks CCD, FgVR. In future, we plan to extend the DB Dataset to a multi-class dataset for the task of FgVR and explore other notions of criticality for the task of CCD.

7. REFERENCES

[1] Yu-Gang Jiang, Chong-Wah Ngo, and Jun Yang, "Towards optimal bag-of-features for object categorization and semantic video retrieval," in *Image and video retrieval*, 2007.

[2] BV Patel and BB Meshram, "Content based video retrieval systems," *arXiv*, 2012.

[3] Florian Schroff, Dmitry Kalenichenko, and James Philbin, "Facenet: A unified embedding for face recognition and clustering," in *CVPR*, 2015.

[4] Marc T Law, Nicolas Thome, and Matthieu Cord, "Learning a distance metric from relative comparisons between quadruplets of images," *International Journal of Computer Vision*, 2017.

[5] Weihua Chen, Xiaotang Chen, Jianguo Zhang, and Kaiqi Huang, "Beyond triplet loss: a deep quadruplet network for person re-identification," in *CVPR*, 2017.

[6] Jian Wang, Feng Zhou, Shilei Wen, Xiao Liu, and Yuanqing Lin, "Deep metric learning with angular loss," in *ICCV*, 2017.

[7] Colin Lea, Austin Reiter, René Vidal, and Gregory D Hager, "Segmental spatiotemporal cnns for fine-grained action segmentation," in *ECCV*, 2016.

[8] Colin Lea, René Vidal, and Gregory D Hager, "Learning convolutional action primitives for fine-grained action recognition," in *ICRA*, 2016.

[9] Colin Lea Michael D Flynn René and Vidal Austin Reiter Gregory D Hager, "Temporal convolutional networks for action segmentation and detection," *arXiv*, 2017.

[10] Sebastian Stein and Stephen J McKenna, "Combining embedded accelerometers with computer vision for recognizing food preparation activities," in *Ubicomp*, 2013.

[11] ZongYuan Ge, Chris McCool, Conrad Sanderson, Peng Wang, Lingqiao Liu, Ian Reid, and Peter Corke, "Exploiting temporal information for dcnn-based fine-grained object classification," in *Digital Image Computing: Techniques and Applications (DICTA)*, 2016.

[12] Yixin Gao, S Swaroop Vedula, Carol E Reiley, Narges Ahmidi, Balakrishnan Varadarajan, Henry C Lin, Lingling Tao, Luca Zappella, Benjamın Béjar, David D Yuh, et al., "Jhu-isi gesture and skill assessment working set (jigsaws): A surgical activity dataset for human motion modeling," in *M2CAI*, 2014.

[13] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012.

[14] "Pyscenedetect," `https://pyscenedetect.readthedocs.io/en/latest/`.

[15] "Shot detect," `https://github.com/johmathe/Shotdetect`.

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

[17] Ahmed Hassanien, Mohamed Elgharib, Ahmed Selim, Mohamed Hefeeda, and Wojciech Matusik, "Large-scale, fast and accurate shot boundary detection through spatio-temporal convolutional neural networks," *arXiv*, 2017.

[18] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *ICCV*, 2015.

[19] Jiang Wang, Thomas Leung, Chuck Rosenberg, Jinbin Wang, James Philbin, Bo Chen, Ying Wu, et al., "Learning fine-grained image similarity with deep ranking," *CVPR*, 2014.

[20] C. Weihua, C. Xiaotang, Z. Jianguo, and H. Kaiqi, "Beyond triplet loss: a deep quadruplet network for person re-identification," in *CVPR*, 2017.