UCF-CRCV at TRECVID 2015: Semantic Indexing

Amir Mazaheri, Mahdi M. Kalayeh, Haroon Idrees, and Mubarak Shah

Center for Research in Computer Vision, University of Central Florida

Abstract. This paper describes the system we used for the main task of Semantic INdexing (SIN) at TRECVID 2015. Our system uses a five-stage processing pipeline including feature extraction, pooling, encoding, classification and reranking. We employed CNN-based representations, as well as Dense and Root SIFTs as features for our system. We also report results of our experiments with SentiBank features and data augmentation techniques that did not contribute to the performance of the final system. Our second run 'Rostam' achieved an infAP of 26.67% on the 30 concepts evaluated for SIN 2015.

1 Introduction

Semantic Indexing is used as an approach for content-based video retrieval. The main task in Semantic Indexing is defined as 'Given the test collection, master shot reference, and single concept definitions, return for each target concept a list of at most 2000 shot IDs from the test collection ranked according to their likelihood of containing the target' [1]. Based on the training data used in the system, each method can be divided into one of the following types:

- Type A: 'used only IACC training data'
- Type B: 'used only non-IACC training data'
- Type C: 'used both IACC and non-IACC TRECVID (S and V and/or Broadcast news) training data'
- Type D: 'used both IACC and non-IACC non-TRECVID training data'

In our training we used both IACC and non-IACC non-TRECVID data (since CNN [2] is trained on ImageNet). Thus, all of our runs are of Type D. The rest of the paper is organized as follows: Section 2 reviews the pipeline which was used for the main task. Section 3 describes the features and descriptors used in our pipeline, whereas Section 4 presents some experiments we did on the SIN 2014 Test set. Finally, we show quantitative performance of our four submissions in the SIN 2015 challenge in Section 5.

2 System Overview

The overview of our system is shown in Figure 1. We used training images available to all the participants [3]. Each image is divided into different spatial regions. Features are extracted and encoded for each region, the representations for different regions are then concatenated, and a classifier is learned for each feature individually. The decision values are fused at the end (late fusion) to obtain the final concept detection scores for each shot. The concept scores for shots in each video are then adjusted using reranking.

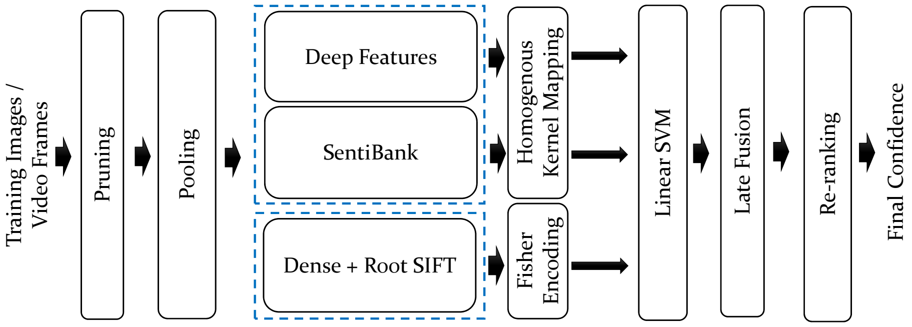

Fig. 1. This figure shows the pipeline of our system. Images are pruned and divided into spatial regions. Different features are extracted and encoded for each region and a SVM is trained for each feature individually. For a given test video, the keyframes in the shot are fused with average pooling, and concept scores for shots are reranked using video information.

3 Features and Descriptors

In our SIN 2015 system, we extracted four different features. Two of them are CNN-based representations, while the other two are Dense SIFT and Root SIFT. We also experimented with SentiBank features, but due to low performance, they were not included in the final system.

- CNN features - Relu$_6$ and FC$_7$: In order to extract CNN features we used the network proposed in [2]. The network is trained on ImageNet training images [4]. We used the representations from two layers: the output of Rectified Linear (Relu) Unit of 6th layer (Relu$_6$), and the output of last fully connected layer (FC$_7$), both of which are 4096 dimensional vectors. We used a total of eight regions (full image, four quadrants and three horizontal slices) for complete representation of each keyframe / image. The final representation is the concatenation of all 8 regions which makes a 32768 dimensional vector, each for Relu$_6$ and FC$_7$. We encoded them with Homogenous Kernel Mapping [5] with $n = 5$ resulting in $2n + 1$ dimensional vectors, which for our case is 360448 $(4096 * 8 * 11)$.

- Dense SIFT and Root SIFT: We densely sampled SIFT points in a given image at multiple scales of $\frac{2}{3}, 1, \frac{4}{3}$ and $\frac{5}{3}$ with a step size of 6 pixels. Each point is then described with SIFT descriptor which is normalized with ℓ_2-norm. In addition to Dense SIFT, we also compute RootSIFT [6] descriptor at each point. For that, we first ℓ_1 normalize the original SIFT descriptor, compute square root of each bin and then normalize the result with ℓ_2-norm. We encoded both descriptors using Fisher vector [7]. We used PCA to reduce the dimensionality of 128d SIFT descriptors to 80 to reduce dimensions and de-correlate the data. We randomly selected about 1 million low-level descriptors from training data and fit a GMM with 256 components. This GMM was later used for aggregating low-level descriptors through Fisher vector framework. Power and ℓ_2 normalizations were applied to compute the Fisher vectors. Our final representation was a $327680(2 * 80 * 256 * 8)$ dimensional vector per image as we used spatial pyramid (1×1, 2×2 and 3×1).
- SentiBank: In one of our recent works [8], the SentiBank [9] detectors boosted the performance of quantifying and predicting the popularity of selfie images. The SentiBank detectors capture mid-level adjective-noun pairs (such as *happy face, colorful clouds...*) depicted in an image. We treated them as mid-level features with a 2089-d vector per image or keyframe. Spatial Pooling was not used for SentiBank features. Similar to CNN-features, we encoded them with Homogenous Kernel Mapping [5] with $n = 5$ resulting in vectors of length 22979 $(2089 * 11)$.

4 Experiments on SIN 2014 Test set

In this section, we report results of some of our experiments which eventually led to the design of final system used in SIN 2015 challenge. All of the results in this section as reported using SIN 2014 Test set.

The first set of experiments involves the optimal configuration and setting of CNN features, summarized in Table 1. The mean fusion of Relu_6 and FC_7 gives better performance than either of them. For all the experiments, we augmented the training data with a horizontally flipped version of the original images. For 'MAX'-pooling, we took the maximum of each bin for the 8 image features, whereas the same features were concatenated for the 'Stacked' variation. Stacking outperforms max-pooling by at least 2% for different combinations of Relu_6 and FC_7.

In Table 2, we show results of different SVM settings for the Dense and Root SIFT as well as SentiBank. LibSVM [10] The first experiment involved selecting optimal value for parameter C. A larger value of $C > 1$ puts more weight on the constraints and resulting in longer training times. From Table 2, we can see that $C = 1$ works best. We also tries Hinge-loss against Squared-Hinge loss (available in VLFeat [11]) as the latter is smoother than the former. Our experiments revealed that Hinge-loss is better for SIN concept detection than Squared-Hinge loss. Furthermore, we speculate that SentiBank features did not work well on SIN

Table 1. This table reports the results of feature representation using Deep Convolutional Neural Networks (Deep CNNs). We used the pre-trained version of AlexNet [2]. These results are reported on the Test set of SIN 2014.

Network	Image Pooling	Augmentation	Layer Number	infAP
7 layers	MAX	Flip	FC7	16.20%
7 layers	MAX	Flip	Relu6	17.67%
7 layers	MAX	Flip	FC7–Relu6 (mean)	20.20%
7 layers	Stacking	Flip	FC7	20.10%
7 layers	Stacking	Flip	Relu6	21.43%
7 layers	Stacking	Flip	FC7–Relu6 (mean)	22.10%

concept detection because of two reasons. First, SentiBank detectors are trained on a clean dataset whereas IACC dataset is extremely noisy and low resolution with most videos taken ‘in the wild’. Second, SentiBank is human-centric as it quantifies various sentiments. On the other hand, videos for SIN competition are not always human-centered and sentiments depicted in such videos do not offer much information for the task of concept detection.

Table 2. This table shows the results of different SVM settings for the three non-CNN features we experimented with when designing our system. SVM parameter $C = 1$ with Hinge-loss works best on average over all the 30 concepts in SIN 2014 Test set.

Feature	LibSVM			VLFeat			
	C=1	C=5	C=10	C=1	C=1	C=5	C=5
				Hinge	Sq-Hinge	Hinge	Sq-Hinge
DSIFT	16	14.28	14.25	15.91	15.51	14.31	14.08
RSIFT	15.52	13.79	13.74	-	-	-	-
SentiBank	8.64	8.05	6.83	8.54	7.94	8.05	7.95

Finally, we show the improvement achieved through reranking. Our approach is similar to [12]. The idea is based on the assumption that the probability of a concept occurring in a shot increases if a similar concept occurs in other shots in a video. Given shots that belong to the same video, we find the maximum score of each concept across all shots in the video, and add a small percentage of the maximum score to the same concept detection scores of different shots. Figure 2 shows the improvement in performance on the 30 concepts evaluated for SIN 2014. On average, reranking gives an improvement of about $2 - 3\%$.

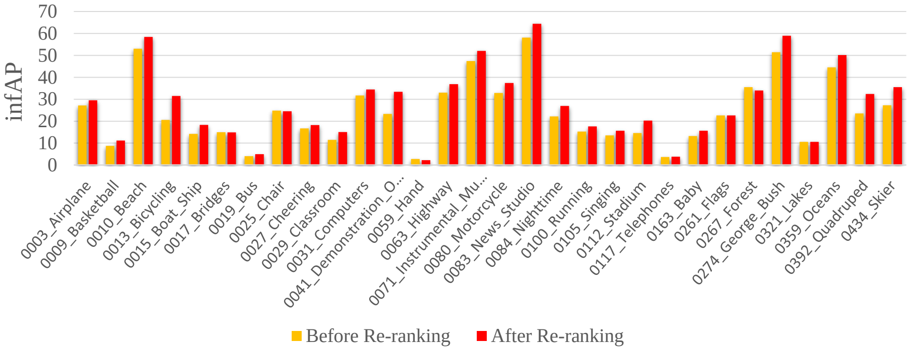

Fig. 2. This graph shows the effect of reranking shots on SIN 2014 Test set. Reranking maintains or improves the performance for 28 out of 30 concepts.

5 SIN 2015 Challenge Results

In this section, we review the performance of four runs we submitted to SIN 2015 challenge.

- Run1-Zaal: For our first run, we used the Relu$_6$, FC$_7$, Dense SIFT and Root SIFT. For classification we used SVM with linear kernel and all the scores were fused using mean fusion. The final score for a shot is found by taking the maximum concept scores across key-frames of that shot (max-pooling). Finally, we performed reranking on the final scores for each concept in each shot.
- Run2-Rostam: Our second run is similar to the first run with the difference that instead of using mean fusion, the linear weights for fusion were learned on SIN 2014 Test set.
- Run3-Tahmine: For this run, we mined hard-negatives and re-trained SVM using hard-negatives as the negative data.
- Run4-Simorgh: This run is the mean fusion of the previous two runs.

In Fig. 3 the results of all four submitted runs are shown. *Run2-Rostam* has the best performance among all the runs. The best results is for concepts *Anchorperson*, *Computers* and *Motorscyle*. The lowest performing concepts are *Animal*, *Boy*, and *Demonstration or Protest*.

Figure 4 shows the best and average infAP reported among all the submission to TRECVID SIN 2015 for each concept. Almost for all the concepts, the infAP obtained by our system is significantly higher than average which proves the effectiveness of the features used in our system. Finally, in Figure 5 we show the ranking of our system compared to other participants (only the top 54 out of 86 submitted runs are shown). With the infAP=26.67%, we are ranked 7^{th} among all the teams which participated in TRECVID SIN 2015.

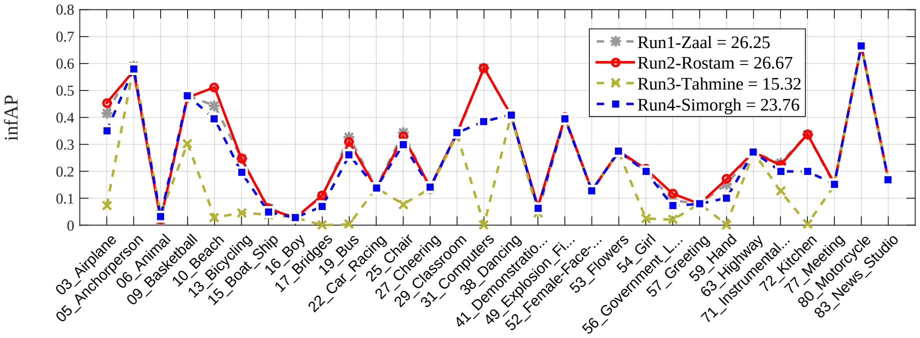

Fig. 3. The average infAP of all four submitted runs: This year only 30 concepts out of 60 concepts were used for evaluation. Run2-Rostam has the best performance (infAP=26.67%) in the SIN 2015 challenge.

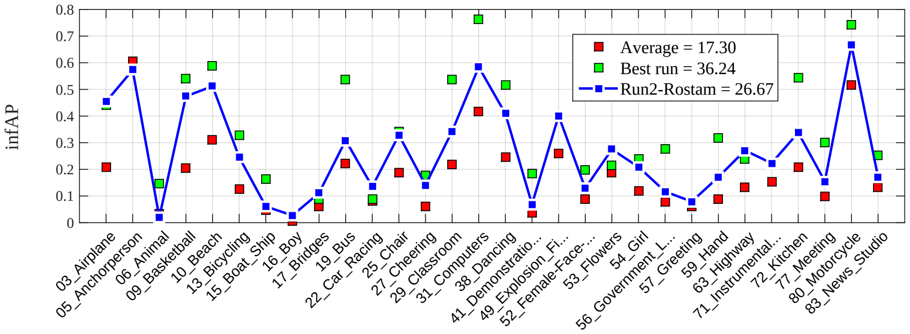

Fig. 4. This figure shows comparison of our method (Run2-Rostam) with mean and maximum infAP reported for each concept for SIN 2015 challenge.

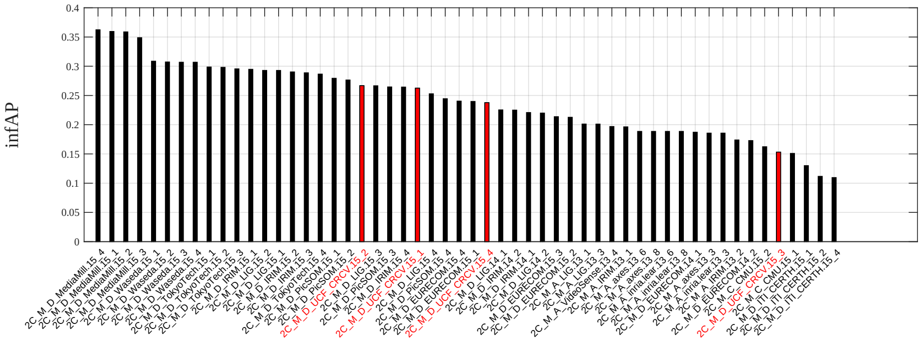

Fig. 5. Top 54 submitted runs to TRECVID SIN 2015. Our runs are the ones shown in orange.

Acknowledgment

This work was supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11PC20066. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the U.S.Government.

References

1. Awad, G., Fiscus, J., Joy, D., Michel, M.: Trecvid 2015 – an overview of the goals, tasks, data, evaluation mechanisms and metrics. In: Proceedings of TRECVID 2015, NIST, USA (2015)
2. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In Pereira, F., Burges, C., Bottou, L., Weinberger, K., eds.: Advances in Neural Information Processing Systems 25. Curran Associates, Inc. (2012) 1097–1105
3. Ayache, S., Qunot, G.: Video corpus annotation using active learning. In: 30th European Conference on Information Retrieval (ECIR'08). (2008)
4. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: Imagenet large scale visual recognition challenge (2014)
5. Vedaldi, A., Zisserman, A.: Efficient additive kernels via explicit feature maps. Pattern Analysis and Machine Intelligence, IEEE Transactions on **34** (2012)
6. Arandjelović, R., Zisserman, A.: Three things everyone should know to improve object retrieval. In: Computer Vision and Pattern Recognition (CVPR) IEEE Conference on. (2012)
7. Perronnin, F., Sánchez, J., Mensink, T.: Improving the fisher kernel for large-scale image classification. In: Computer Vision–ECCV 2010. Springer (2010) 143–156
8. Kalayeh, M.M., Seifu, M., LaLanne, W., Shah, M.: How to take a good selfie? In: Proceedings of the 23rd Annual ACM Conference on Multimedia Conference. (2015) 923–926
9. Borth, D., Chen, T., Ji, R., Chang, S.F.: Sentibank: large-scale ontology and classifiers for detecting sentiment and emotions in visual content. In: Proceedings of the 21st ACM international conference on Multimedia, ACM (2013) 459–460
10. Chang, C.C., Lin, C.J.: Libsvm: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology (TIST) **2** (2011) 27
11. Vedaldi, A., Fulkerson, B.: Vlfeat: An open and portable library of computer vision algorithms. In: Proceedings of the international conference on Multimedia, ACM (2010) 1469–1472
12. Inoue, N., Shinoda, K.: n-gram models for video semantic indexing. In: Proceedings of the ACM International Conference on Multimedia. (2014) 777–780