DLDR: Deep Linear Discriminative Retrieval for cultural event classification from a single image

Rasmus Rothe, Radu Timofte, Luc Van Gool
Computer Vision Lab, D-ITET, ETH Zurich, Switzerland
{rrothe,timofter,vangool}@vision.ee.ethz.ch

Abstract

In this paper we tackle the classification of cultural events from a single image with a deep learning based method. We use convolutional neural networks (CNNs) with VGG-16 architecture [17], pretrained on ImageNet or the Places205 dataset for image classification, and fine-tuned on cultural events data. CNN features are robustly extracted at 4 different layers in each image. At each layer Linear Discriminant Analysis (LDA) is employed for discriminative dimensionality reduction. An image is represented by the concatenated LDA-projected features from all layers or by the concatenation of CNN pooled features at each layer. The classification is then performed through the Iterative Nearest Neighbors-based Classifier (INNC) [20]. Classification scores are obtained for different image representation setups at train and test. The average of the scores is the output of our deep linear discriminative retrieval (DLDR) system. With 0.80 mean average precision (mAP) DLDR is a top entry for the ChaLearn LAP 2015 cultural event recognition challenge.

1. Introduction

Image classification is at the core of computer vision. Extensive literature is devoted to the study of classification of images into objects and/or scenes. The recent advances due to the introduction of large datasets such as ImageNet [15] or PASCAL VOC [5] for object classification and the use of deep learning techniques [3, 7, 11] brought into focus the classification of more demanding classes such as 'cultural events' where the geometry and/or appearance of a single object or scene are not anymore the dominant features determining the class. Particularly, a picture of a cultural event depends entirely on the photographer's subjectivity. Each such picture is just a narrow view of what happens under the big umbrella of the cultural event. Classification and retrieval of images of cultural events are of interest for many people, especially tourists. There are multiple important cultural events in the world that attract lots of participants and produce huge amounts of photos to browse.

Figure 1. Cultural event images and class labels from LAP dataset.

In this paper we tackle the classification of cultural events from a single image, a consumer photograph, with a deep learning-based method and report our performance on the cultural event recognition dataset of the ChaLearn Looking at People 2015 (LAP) challenge [4] (see Fig. 1).

We use convolutional neural networks (CNNs) with VGG-16 architecture [17], pretrained on the ImageNet dataset [15] or the Places205 dataset [25] for image classification, and fine-tuned on cultural events training data from LAP. Our CNN features are the fully-connected (fc) layer 7 with 4096 dimensions. We follow a layered approach (see Fig. 2). For each layer, CNN features are robustly extracted from each image over a grid. At each layer, Linear Discriminant Analysis (LDA) [6] is employed for reducing the dimensionality of the CNN features and to embed discriminativity. An image is represented by the concatenated LDA-projected features from all layers or by the concatenation of the average pooled raw CNN features at each layer. The classification is handled through the Iterative Nearest Neighbors-based Classifier (INNC) [20, 21]. Classification scores are obtained for different image representation setups at train and test. The average of the scores is the output of our deep linear discriminative retrieval (DLDR) system.

DLDR is a top entry for the ChaLearn LAP 2015 cultural event recognition challenge with 0.80 mean average precision (mAP), 0.05 below the best reported result.

Next we review work related to our task and method. Section 2 introduces our DLDR method. Section 3 describes the experiments and discusses the results, while in Section 4 we conclude the paper.

1.1. Related work

The ChaLearn Looking at People challenge on cultural event recognition from single images in conjunction with CVPR 2015 [1] is the precursor of the ChaLearn LAP challenge in conjunction with ICCV 2015 [4] that we targeted in this paper. The previous challenge used a 50 classes dataset while the new one extended it by proposing a larger dataset with 100 classes. The solutions proposed for the previous challenge are those most related to our own. In Table 5 are the top 4 ranked teams of that challenge. Next, we present them in relation to our proposed DLDR method.

MMLAB: The solution of Wang *et al*. [23] fuses five types of CNNs. These are ClarifaiNet [24] pretrained on the ImageNet dataset, AlexNet [11] pretrained on the Places205 dataset, GoogleNet [18] pretrained on the ImageNet dataset and the Places205 dataset, and VGG-19 [17] pretrained on the ImageNet dataset. All of them are fine-tuned on the cultural event training data and the scores are fused by weighting for the final results. MMLAB ranked 1^{st} with 0.85 mAP, significantly more than the next team with 0.76 mAP. Our DLDR is significantly lighter, it uses only one kind of CNN, the VGG-16, pretrained on ImageNet and on Places205. DLDR also fine-tunes and fuses scores for the final results, but in addition uses multiple layers in the representations, discriminant projections, and INNC classifiers.

UPC-STP: The team of Salvador *et al*. [16] combines features from the fully connected (fc) layers of a CNN pretrained with ImageNet and a second one fine-tuned on the cultural event dataset. For each fc layer, Linear SVMs are trained for the corresponding features. These are further fused using an SVM. A temporal model of the events is learned and used to refine the outputs. Our DLDR uses another CNN architecture, pretrains also on ImageNet, uses only the last fc layer as CNN raw features and employs a different classification strategy.

MIPAL_SNU: The team of Park and Kwak [14] assumes that the discriminant image regions are the ones relevant to classification. Therefore, they first extract meaningful image regions of various size. Then they train a CNN with 3 convolutional layers and pooling layers, and 2 fc layers. The probability distribution for the classes is calculated for every image region selected from the test image and class probabilities are computed.

SBU_CS: The team of Kwon *et al*. [12] studies SIFT, SIFT+color, and CNN features in combination with 3 layer spatial pyramid matching (SPM) [13] and a regularized max pooling (RMP) [9] technique. The CNN is pretrained on ImageNet and no fine-tuning is employed. Their best combination is a SPM with SIFT+Color and RMP with CNN features. Our DLDR method also uses layered representations and CNN features.

The novelty of our proposed method lies in using LDA discriminative projections of CNN features at different pyramidal layers and per layer pooled CNN features to improve classification accuracy. Furthermore, we extend the formulation of the INNC classifier with weight-spreading to better deal with retrieval of a large number of classes.

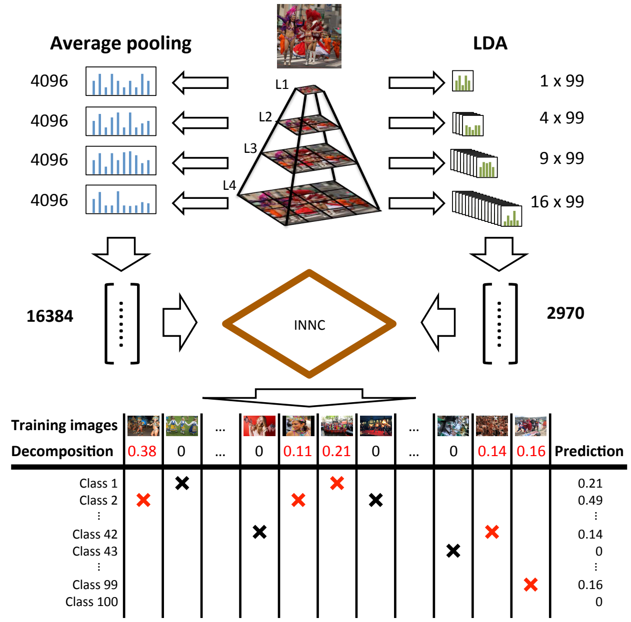

Figure 2. Pipeline for our DLDR method.

2. Proposed method (DLDR)

In this section we describe the proposed method: deep linear discriminative retrieval (DLDR). The scheme of DLDR is shown in Fig. 2.

2.1. Deep learning

Our DLDR is based on the deep learned representations of image regions. We employ CNNs with the VGG-16 [17] architecture which provides a good balance between the representation power and time plus memory requirements. Simonyan *et al*. [17] achieve state-of-the-art results with this architecture on benchmarks such as ImageNet [15].

2.1.1 Pretraining

Without a (very) large training set of images getting trained from scratch a CNN with a very large number of parameters

Figure 3. Examples of images where DLDR is successful in a top-1 evaluation.

Figure 4. Examples of images where DLDR fails in a top-1 evaluation.

(like ours) is cumbersome and likely to overfit and to produce poor results. Therefore, for cultural events recognition we use as a starting point the CNN pretrained on the ImageNet dataset [17] and CNN pretrained on the Places205 dataset [22]. Previously Wang *et al*. [23] also used these two datasets for pretraining models for cultural event recognition.

2.1.2 Training

We train two separate CNNs on the provided LAP dataset corresponding to those pretrained on ImageNet and Places205, resp. We adapt the output layer of the network to have a number of neurons equal to the number of classes, here 99 cultural events and a 'Non-Class' as in the LAP dataset. The training data, consisting of the provided LAP training dataset and LAP validation dataset, was randomly split into 90% used for training and 10% for testing. We kept the distribution of classes the same in both sets. Our training set is further enlarged by augmentation. 10 random crops from each original training image are added to the training set. Each random crop has at least half the side length of the original image.

2.2. Layered representations

Inspired by [8, 13], we extract CNN features in a pyramidal fashion. Specifically we extract features at 4 scales. In the first level we extract features over the entire image, in the second, third, and fourth level we extract from 2×2, 3×3, and 4×4 regions, resp. The regions overlap with 50%. We scale each image region to 256×256 and then extract the last feature layer (fc7, 4096 dimensions) for 10 different crops at a size of 224×224 in each corner and the center of the image, as in [7]. We do the same for the flipped version of the image. The features of these 10 crops are then averaged to give the final feature representation. This results in $1^2 + 2^2 + 3^2 + 4^2 = 30$ feature representations of 4096 dimensions.

We can not handle representations of 30 concatenated raw CNN features. We either employ encoding over a visual codebook as in standard SPMs (we got discouraging preliminary results), reduce the dimensionality (i.e. through LDA), or pool the raw features at each layer.

2.2.1 Pooled CNN features

The idea of pooling directly the raw CNN features without caring about their image positions is inspired by the robust prediction commonly employed by CNN solutions (predicting on different crops around the desired image region of interest). We considered different pooling operators and found average pooling to be the best in robustness and performance. Our pooled features are average pooled raw CNN features at a given layer. Correspondingly, the layered representation, called R1, has $n \times 4096$ dimensions where n is the number of layers with pooled raw CNN features. In our case the representation is high dimensional ($4 \times 4096 = 16384$ dimensions) and thus can capture subtle details learned by the CNN.

2.2.2 LDA-projected features

Due to limited computational resources we explored efficient dimensionality reduction methods. Principal component analysis (PCA), a natural choice, loses quite a bit of performance even for reduction factors of 2 or 4. Since reducing the dimensionality while preserving the energy is challenging, we picked linear discriminant projections that would thus compensate for the loss in dimensions by improving the discriminative power.

Linear Discriminant Analysis (LDA) maximizes the ratio of the between-class scatter and the within-class scatter. We use LDA in its regularized form [6], with regularization parameter set to 1, as implemented by Cai *et al*. [2]. In our preliminary experiments, LDA and its 99-dimensional projections (number of classes - 1) were able to provide for equal and better classification performance than the original raw features, while SRLP [19] (which embeds sparse relations) needs 200 dimensions to improve over the LDA-projections.

Table 1. mAP (%) on our validation set (2863 of 20036 images) for different configurations.

Layers	Encoding	CNN pretrained on ImageNet	CNN pretrained on Places205	Fusion
L1	Raw	74.59	73.61	77.32
	LDA	73.91	73.96	77.64
L2	Raw	76.16	73.70	77.90
	LDA	77.90	75.92	79.69
L3	Raw	75.43	72.20	76.54
	LDA	77.65	75.39	79.03
L4	Raw	73.63	69.00	74.18
	LDA	77.28	73.14	77.73
L1+L2	Raw	76.75	75.16	78.75
	LDA	78.00	76.62	80.05
L1+L2+L3	Raw	77.52	75.80	79.24
	LDA	79.00	77.12	80.22
L1+L2+L3+L4	Raw	77.63	75.84	79.25
	LDA	79.10	76.93	80.12

Table 2. mAP (%) of DLDR on our validation set (2863 of 20036 images).

Train/test representation	ImageNet	Places205	Fusion
C1	77.63	75.84	79.25
C2	79.10	76.93	80.12
C3	79.26	76.77	80.16
C4	79.36	77.08	80.38
C2+C3+C4	79.61	77.29	80.47
C1+C2	79.56	77.56	80.46
C1+C2+C3+C4	**79.96**	**77.74**	**80.70**

Table 3. Classification on our validation set (2863 of 20036 images)

	ImageNet	Places205	Fusion
Linear SVM	77.04	75.58	78.87
INNC	78.42	76.15	79.76
INNC-KNN	79.10	76.93	80.12

We learn a separate Linear Discriminant Analysis (LDA) projection for each of the 4 layers in our representation. We then concatenate the LDA-projected features to form a feature vector of $30 \times 99 = 2970$ dimensions, representation R2. Additionally we construct a flipped representation of R2 by horizontally flipping the local representation for the 2^{nd}, 3^{rd} and 4^{th} layer, called R3. Note that R3 is a permutation of the features of R2.

The LDA helps to not only reduce the dimensionality but also to embed discriminativeness into the features.

2.3. Classification

For classification we use the Iterative Nearest Neighbors-based Classifier (INNC) of Timofte and Van Gool [20]. The INN representation [21] is the result of a sparse linear decomposition of the query sample over the training pool. The weights belong to $[0, 1)$ and sum up to 1. For each class, the weights corresponding to training samples of that class are summed up. The class with the largest impact in the INN decomposition of a query is the INNC prediction. We set the maximum number of non-zeros (neighbors) to $K = 14$ and the regularization parameter to $\lambda = 0.1$. For each test sample we obtain an INN representation over the training set. This sparse matrix of weights is then used for classification. The probability for a given test sample to belong to a class is taken as the sum of the weights corresponding to all training samples of that class. As the INN representation is sparse ($\leq K$), often with fewer non-zero weights than classes, many classes have a probability of 0. To overcome this issue we extend the formulation of INNC by additionally spreading the weights also to the nearest neighbors of the training samples, with some exponential decay (0.75^r, where r is the rank of the neighbor). This helps to increase retrieval performance especially on difficult samples.

INNC is applied to the representations separately:

C1: R1

C2: R2 in the training set and R2 at testing

C3: R3 in the training set and R2 at testing

C4: R2 and R3 in the training set and R2 at testing

Note that if we would have had R2 in the training set and R3 at testing, this would be the same as C3 as R2 and R3 just differ by permutation. We obtain those predictions for both networks, resulting in 8 predictions in total which are averaged fused to give the final DLDR prediction score.

3. Experiments

3.1. Dataset and evaluation protocol

The ChaLearn LAP cultural event recognition dataset [4] consists of 28705 images collected from two images search engines (Google Images and Bing Images). The images contain photos from 99 important cultural events around the world and a non-class. The dataset is split into three parts, 50% for training (14332 images), 20% for validation (5704 images), 30% for testing (8669 images). There are approximately the same number of images in each class with the exceptions of the non-class having around ten times as many images.

In our paper the results are evaluated as defined for the ChaLearn LAP challenge. Specifically, for a given class the average precision (AP) is calculated by measuring the area

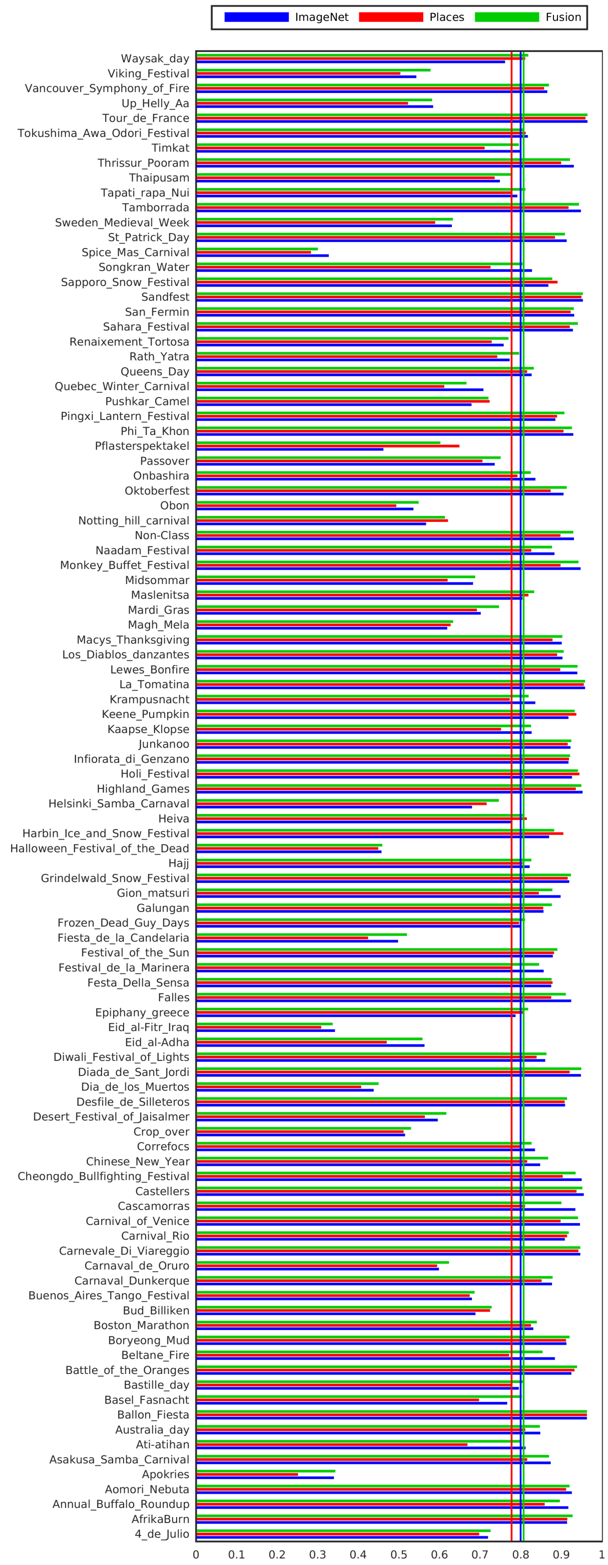

Figure 6. DLDR average precisions (AP) for LAP classes using Places205 pretraining, ImageNet pretraining, or the fused predictions.

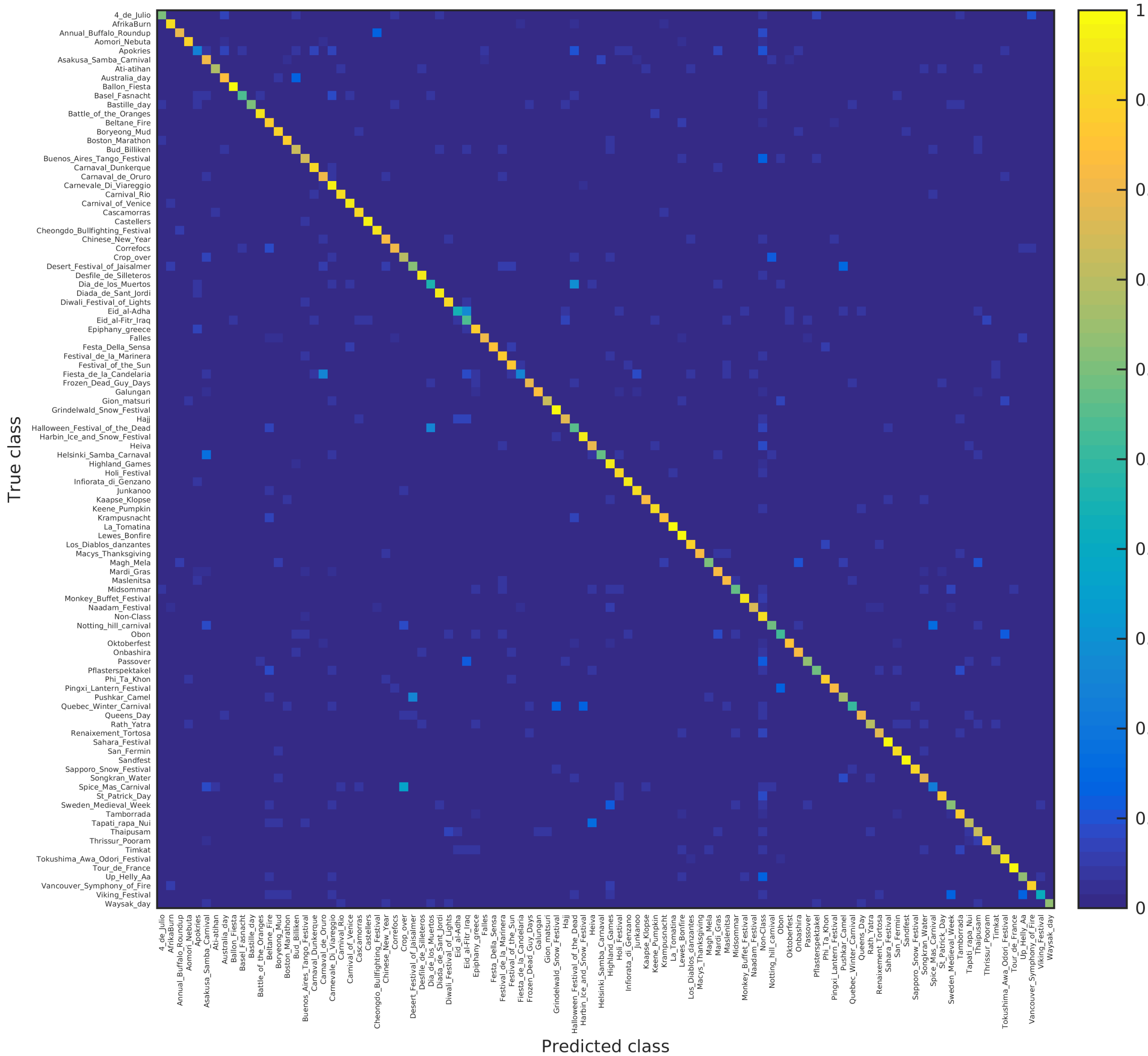

Figure 5. Confusion matrix for our DLDR system on the LAP classes. Best seen on screen.

under the precision/recall curve. The AP scores are then averaged over all 100 classes to form the final mean average precision (mAP).

3.2. Implementation details

Our DLDR pipeline is written in Matlab. The CNNs are trained on Nvidia Tesla K40C GPUs using the Caffe framework [10]. The machine used for calculating the LDA projections and classification has 128 GB of memory.

Training each of the two CNNs took about 30 hours. At test time extracting the features over all training and testing images over all 4 layers ($30 \times 10 = 300$ extractions per image) took around 100 hours. Calculating the LDA projections and classification took around 3 hours.

The source codes are publicly available at: http://www.vision.ee.ethz.ch/~timofter

3.3. Validation results

We compare the performance of our proposed method for different feature representations, layers and pretrained networks. In Table 1 the results are summarized. The performance is shown for the case where features are extracted only at one layer, as well as for the case when features with increasing depth are combined. For both the pooled CNN features as well as the LDA-encoded features the performance increases from L1 to L2. Beyond L2, for L3 and

L4 performance decreases again suggesting that the level of detail gained from the smaller image regions cannot compensate the insights missing from a global feature representation. Nonetheless, when combining the features from various layers, performance increases as one adds more layers to the representation. This suggests that the increase in dimensionality of the feature vector is well compensated by the extra insights from the smaller image regions. Comparing the pooled CNN features with the LDA-projected features, in the majority of the experimental setups, the LDA features perform around 2% better. The weights after pretraining on ImageNet give consistently better results than the Places205 weights. However, the fusion of the two then again improves performance by at least 1%. Overall we are able to push the performance from 74.59% to 80.12% when combining the LDA-projected features from all layers and the ImageNet and the Places205 pretraining compared to when just extracting pooled features from ImageNet on the entire image (L1).

For classification we compare the performance of our proposed INNC with weight spreading (INNC-KNN) to the conventional INNC and Linear SVM as shown in Table 3. When fusing the features from ImageNet and the Places205 dataset INNC improves 1% over Linear SVM. Weight spreading further improves performance by 0.5%

Table 2 shows the performance when combining the 4 different classifications for each network, resulting in 8 predictions in total. Combining the LDA-projected predictions with its flipped version (C2+C3+C4) improves performance by around 0.5%. Also combining the LDA-projected features with the pooled CNN features gives an improvement of 0.5% over the LDA features and more than 1% over the pooled features. Combining all 8 predictions then leads to an overall improvement of 1.5% over just using the pooled CNN features. Overall this improves the performance of just using R1 by 1.5% up to 80.70% on our validation set.

In Figure 6 we compare the performance of the pretrained ImageNet network and the Places205 network. For the majority of the classes the fusion of the two networks outperforms the individual networks. Pretraining on ImageNet generally gives better results than when the network was pretrained on the Places205 dateset. However, there are some exceptions, i.e. for the classes Pflasterspektakel, Waysak_day, and Pushkar_Camel the pretrained Places205 network seems to give better accuracy.

As some of the classes are very similar, i.e. there are multiple carnival events, we investigated the confusion between classes. Specifically we assigned each image to the class with the largest confidence and then visualized the inter-class confusion (see Fig. 5). Some classes like Eid_al-Adha and Eid_al-Fitr_Iraq or Pushkar_Camel and Desert_Festival_of_Jaisalmer have a high confusion, which is also confirmed when looking at the images from the classes – as a human it is nearly impossible to distinguish between them.

Table 4. ChaLearn LAP 2015 final ranking on the test set. 67 registered participants.

Rank	Team	mAP
1	VIPL-ICT-CAS	0.85
2	FV	0.85
3	MMLAB	0.84
4	NU&C	0.82
5	**CVL_ETHZ (ours)**	**0.80**
6	SSTK	0.77
7	MIPAL_SNU	0.76
8	ESB	0.76
9	Sungbin Choi	0.62
10	UPC-STP	0.58

In Figure 3 we visualize cases where our proposed method successfully recognizes the correct class. The system seems to successfully pick up subtle details which are typical for the event (i.e. the US flag for 4th of July or the floral wreath for Midsommar).

Figure 4 shows some failure cases. In many of those cases the classes are either very similar (i.e. same type of event, same location, same vegetation) or the image shows just one large object and it is thus hard to directly assign it to a specific class (i.e. just a person, a boat or a building).

3.4. Looking At People (LAP) challenge

The ChaLearn Looking at people (LAP) challenge on cultural event recognition had two phases.

In the first phase, the training and validation images of the LAP dataset were provided to the registered participants. If the training images had class labels, the labels for validation images were unknown until the second phase. For the performance score (mAP) on the validation set each team submitted their results to the server. After the validation phase, the labels for the validation images were released together with the test images. Again, the teams were invited to submit their results on the test images to the competition server without getting to know their performance or rank. The organizers announced the final ranking and scores after the second phase ended. Table 4 shows the final ranking of the ChaLearn LAP challenge on cultural event recognition based on the test set. Our DLDR method ranks 5^{th} with a mAP of 0.80, being only 0.05 below the best reported performance of the VIPL-ICT-CAS team.

This ChaLearn LAP challenge with 100 classes was preceded by a ChaLearn LAP challenge in conjunction with CVPR 2015 which had 50 classes [1]. Most of the top ranked teams, unlike us, participated also in the previous challenge. The top 4 teams in the previous (easier) challenge are listed in Table 5 and their solutions were discussed in the related work section 1.1.

Table 5. CVPR ChaLearn LAP 2015 top 4 ranked teams [1]

Rank	Team	mAP
1	MMLAB [23]	0.85
2	UPC-STP [16]	0.76
3	MIPAL_SNU [14]	0.73
4	SBU_CS [12]	0.61

4. Conclusions

We proposed an effective method for cultural event recognition from single images called Deep Linear Discriminative Retrieval (DLDR). DLDR employs CNNs pretrained on ImageNet and Places205 datasets, and fine-tuned on cultural events data. CNN features are robustly extracted at 4 different layers in each image. They are either average pooled or LDA projected at each layer. Thus, an image is represented by the concatenated LDA-projected features from all layers or by the concatenation of CNN pooled features at each layer. Using our Iterative Nearest Neighbors-based Classifier (INNC), scores are obtained for different image representation setups. The average scores are the fused DLDR output. With 0.80 mean average precision (mAP) our DLDR solution is a top entry in the ChaLearn LAP 2015 cultural event recognition challenge.

Acknowledgements. This work was supported by the ERC Advanced Grant VarCity (#273940).

References

[1] X. Baro, J. Gonzalez, J. Fabian, M. A. Bautista, M. Oliu, H. Jair Escalante, I. Guyon, and S. Escalera. Chalearn looking at people 2015 challenges: Action spotting and cultural event recognition. In *CVPR, ChaLearn Looking at People workshop*, June 2015. 2, 7, 8

[2] D. Cai, X. He, and J. Han. Srda: An efficient algorithm for large-scale discriminant analysis. *Transactions on Knowledge and Data Engineering*, 20(1):1–12, 2008. 4

[3] D. Cireşan, U. Meier, and J. Schmidhuber. Multi-column deep neural networks for image classification. *CVPR*, 2012. 1

[4] S. Escalera, J. Fabian, P. Pardo, X. Baro, J. Gonzalez, H. J. Escalante, and I. Guyon. Chalearn 2015 apparent age and cultural event recognition: datasets and results. In *ICCV, ChaLearn Looking at People workshop*, 2015. 1, 2, 5

[5] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 88(2):303–338, 2010. 1

[6] J. H. Friedman. Regularized discriminant analysis. *Journal of the American statistical association*, 84(405):165–175, 1989. 1, 4

[7] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014. 1, 4

[8] K. He, X. Zhang, S. Ren, and J. Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. In *ECCV*. 2014. 4

[9] M. Hoai. Regularized max pooling for image categorization. In *BMVC*, 2014. 2

[10] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *ACM International Conference on Multimedia*, 2014. 6

[11] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1097–1105, 2012. 1, 2

[12] H. Kwon, K. Yun, M. Hoai, and D. Samaras. Recognizing cultural events in images: A study of image categorization models. In *CVPR, ChaLearn Looking at People workshop*, June 2015. 2, 8

[13] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006. 2, 4

[14] S. Park and N. Kwak. Cultural event recognition by subregion classification with convolutional neural network. In *CVPR, ChaLearn Looking at People workshop*, June 2015. 2, 8

[15] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, pages 1–42, 2014. 1, 2

[16] A. Salvador, M. Zeppelzauer, D. Manchon-Vizuete, A. Calafell, and X. Giro-i Nieto. Cultural event recognition with visual convnets and temporal models. In *CVPR, ChaLearn Looking at People workshop*, 2015. 2, 8

[17] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014. 1, 2, 3

[18] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *CVPR*, 2015. 2

[19] R. Timofte and L. Van Gool. Sparse representation based projections. In *BMVC*, pages 61–1, 2011. 4

[20] R. Timofte and L. Van Gool. Iterative nearest neighbors for classification and dimensionality reduction. In *CVPR*, 2012. 1, 5

[21] R. Timofte and L. Van Gool. Iterative nearest neighbors. *Pattern Recognition*, 48(1):60 – 72, 2015. 1, 5

[22] L. Wang, S. Guo, W. Huang, and Y. Qiao. Places205-vggnet models for scene recognition. *arXiv:1508.01667*, 2015. 3

[23] L. Wang, Z. Wang, W. Du, and Y. Qiao. Object-scene convolutional neural networks for event recognition in images. In *CVPR, ChaLearn Looking at People workshop*, June 2015. 2, 3, 8

[24] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *ECCV*. 2014. 2

[25] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning deep features for scene recognition using places database. In *NIPS*, pages 487–495, 2014. 1