1 Team details

- Team name: FV
- Team leader name: Xiu-Shen Wei
- Team leader address: Nanjing University (Xianlin Campus), 163 Xianlin Avenue, Qixia District, Nanjing 210023, China
- Team leader phone number and email: +86-182-0518-8066, weixs.gm@gmail.com or weixs@lamda.nju.edu.cn
- Rest of the team members: Bin-Bin Gao and Jianxin Wu
- Team website URL (if any): `http://lamda.nju.edu.cn/weixs/`, `http://lamda.nju.edu.cn/gaobb/` and `http://cs.nju.edu.cn/wujx/`.
- Affiliation: National Key Laboratory for Novel Software Technology, Nanjing University

2 Contribution details

- Title of the contribution: Deep Spatial Pyramid and Multiple CNNs Ensemble
- Final score: We can achieve 0.835 mAP on the validation set.
- General method description: The main approach used in our experiments is named as DSP (Deep Spatial Pyramid) [1]. By considering utilizing the spatial information with fully convolutional activations, we form a natural deep spatial pyramid [2] by partitioning an image into sub-regions and computing local features inside each sub-region. In practice, we just need to spatially partition the cells of activations in the last convolutional layer, and then pool deep descriptors in each region separately using Fisher Vector [3]. The operation of DSP is illustrated in Fig. 1. Moreover, in order to capture variations of the activations caused by variations of objects in an image, we generate a multiple scale pyramid, extracted from S (in our experiments, we employed four scales, i.e., 1.4, 1.2, 1, 0.8) different rescaled versions of the original input image. We feed images of all different scales into a pre-trained CNN model and extract deep activations. In each scale, the corresponding rescaled image is encoded into a $2mdK$-dimensional vector by DSP. And then, these S vectors will be merged into a single vector by average pooling. In addition, for each original image, we extract three crops (384×384) from it and flip the original image horizontally as data augmentation. However, because there is one "non-class" in the cultural event recognition data set, we just do the operations of data augmentation on the other 99 cultural events classes, which on one hand can supply diverse data sources, and on the other hand can solve the class

imbalance problem. In this case, one test image will be represented by five equally instances. At the test stage, we average the prediction scores of these five instances to get the final score for this test image. In our experiments, three popular CNN pre-trained models (i.e., VGG16, VGG19 and Place-CNN) were employed to extract the DSP features. Meanwhile, for VGG16 and VGG19, we also fine-tune them on the training and validation images and crops. Therefore, for one image/crop, we can get five DSP features from five CNN models. Then, we concatenate five DSP features into one single vector to represent this image/crop. Finally, we feed these vectors into a logistic regression to build a classifier and use the softmax as the prediction scores for each image/crop.

- References:

  1. B.-B. Gao, X.-S. Wei, J. Wu and W. Lin. Deep spatial pyramid: The devil is once again in the details. arXiv:1504.05277.
  2. S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. CVPR, 2006.
  3. J.Sánchez, F.Perronnin, T.Mensink, and J.Verbeek. Image classification with the fisher vector: Theory and practice. IJCV, 105(3):222–245, 2013.

- Representative image / diagram of the method:

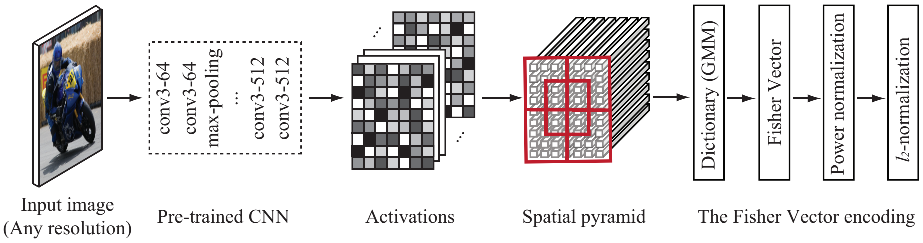

Figure 1: The image classification framework. DSP feeds an arbitrary resolution input image into a pre-trained CNN model to extract deep activations. A GMM visual dictionary is trained based on the deep descriptions from training images. Then, a spatial pyramid partitions the deep activations of an image into m blocks in N pyramid levels. In this way, each block activations are represented as a single vector by the improved Fisher Vector. Finally, we concatenate the m single vectors to form a $2mdK$-dimentional feature vector as the final image representation.

3 Data Preprocessing

- Describe features used or data representation model (if any):

1. Place-CNN: CNN trained on 205 scene categories of Places Database with 2.5 million images. The architecture is the same as Caffe reference network. This model is available at: `http://places.csail.mit.edu/`.
2. VGG16 and VGG19: The models are the improved versions of the models used by the VGG team in the ILSVRC-2014 competition. The details can be found from `http://www.robots.ox.ac.uk/~vgg/research/very_deep/`.

- Dimensionality reduction technique applied (if any): NULL
- Segmentation strategy used (if any): NULL
- Other techniques/strategy used not included in previous items FOR DATA PREPROCESSING (if any): For each original image, we extracted three crops (384 × 384) from it and flipped the original image horizontally as data augmentation.

4 Classification details

- Classifier or method used to train and validate your results (if any): Logistic regression in LIBLINEAR.
- Large scale strategy (if any): We modified the LIBLINEAR package such that it can handle the data scale in this challenge.
- Compositional model used (scene context representation), i.e. pictorial structure (if any): NULL
- Other technique/strategy used not included in previous items FOR CLASSIFICATION (if any): NULL

5 Global Method Description

- Total method complexity analysis: The system can be mostly automatic, and implemented in MATLAB in about 1800 lines. The running time complexity and memory usage complexity can be found in the "Other details" section.
- Which pre-trained or external methods have been used (for any stage, if any): Three pre-trained CNN models have been used. The VGG16 and VGG19 models are pre-trained on the ImageNet dataset, while the Place-CNN model is pre-trained using the scene-centric dataset called Places. Images in these datasets are different from the cultural event images in this challenge.
- Qualitative advantages of the proposed solution:

1. It can handle any resolution images as inputs;
2. It utilized the spatial information with fully convolutional activations;
3. It could capture variations of the activations caused by variations of objects in an image.
4. It can ensemble multiple deep CNNs.

- Results of the comparison to other approaches (if any): Please refer to the [1] reference.

- Novelty degree of the solution and if is has been previously published: The system is built on DSP, a novel system for image categorization. DSP builds universal image representations from CNN models, while adapting this universal representation to different image domains in different applications. DSP is not published, but has a preprint version in the arXiv server.

6 Other details

- Language and implementation details (including platform, memory, parallelization requirements): We mainly used MATLAB to implement our method. Our experiments were executed by MATLAB2014a in the RedHat6.5 system with 99GB memory and four K40 Gpus. In addition, our experiments will produce the training and testing features which will cost about 500GB on disks.

- Human effort required for implementation, training and validation?: Beyond compilation of various source codes and move trained models to appropriate directories in the disk, there is no manual effort required at all.

- Training/testing expended time?: For fine-tune, it will cost about 1∼2 days for each model of VGG16 and VGG19. And for extracting DSP features, it will cost 30 hours for VGG models and 8 hours for Place-CNN. For training a logistic regression, it will cost about 8 hours for 100 classes. But it is very efficient for testing.

- General comments and impressions of the challenge?: Thank you for your efforts and providing a wonderful opportunity for competition and communication.