Cultural Event Recognition by Subregion Classification with Convolutional Neural Network

Sungheon Park and Nojun Kwak
Graduate School of CST, Seoul National University
Seoul, Korea
{sungheonpark,nojunk}@snu.ac.kr

Abstract

In this paper, a novel cultural event classification algorithm based on convolutional neural networks is proposed. The proposed method firstly extracts regions that contain meaningful information. Then, convolutional neural networks are trained to classify the extracted regions. The final classification of a scene is performed by combining the classification results of each extracted region of the scene probabilistically. Compared to the state-of-the-art methods for classifying Chalearn Looking at People cultural event recognition database, the proposed methods shows competitive results.

1. Introduction

Image classification or scene classification is one of the basic tasks in computer vision, and they have drew a lot of attention in recent years. In particular, performance of object detection and classification has been increased significantly in accordance with recent advances on deep learning. While most datasets for object detection or classification such as ImageNet challenge [14] focus on dealing with different kinds of objects, it is also important to classify scenes which contain similar objects common in different scenes. Especially, recognizing and classifying scenes which include people is useful in order to understand cultural characteristics or human behaviors. Chalearn Cultural event classification challenge [1] aims to understand scenes that can be seen at various cultural events throughout the world. The dataset consists of 50 cultural events around the world with about 11,000 images, most of which contain people. Since many images contain similar objects, it is important to find discriminative characteristics of cultural events, for example, garments or makeup of people at the events. It can be expected that these discriminant characteristics will only appear in specific regions of images and that the rest non-specific areas of the images appear commonly in multiple events. Following this rationale, we propose a method that classifies an image by gathering subregion classification results. Figure 1 illustrates the classification process of our method. Rather than classifying the whole input image, we detect regions of the image with different sizes and classify those regions individually using convolutional neural networks (CNN). In our experiments, the CNNs are trained successfully without exploiting pre-

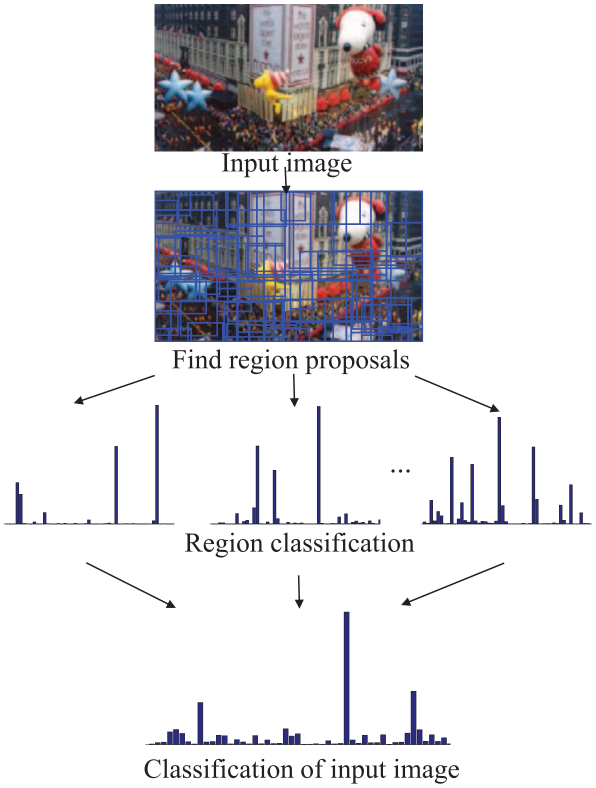


Figure 1: Overview of the proposed algorithm. After finding region proposals from an image, each region proposal is classified by CNN. Then, the final classification of an image is calculated by combining the region classification results.

trained weights. Then, the class probabilities of each region are gathered together to produce the final output for the query image. The proposed framework can be applied not only to the cultural event recognition, but also to any other scene classification problems. Despite the simplicity of our approach, the performance of the proposed algorithm is competitive with those of the top-tier methods in the challenge.

2. Related work

Recent advances in deep neural networks boosted the performance of image classification problems. Since the work of Krizhevsky et al. [10], CNN has received wide attention, and many different structures are proposed to improve classification accuracy [7, 17, 18]. Classification using CNN can also be applied to object detection. Unlike classification of an image, object detection needs to locate the object in the input image. To locate an object in an image, classification algorithm should be applied to all the possible subregions of the image. Girshick et al. [6] reduced the search space of the subregions by extracting region proposals from images. The idea of region proposal is adopted in our framework for a different purpose.

Cultural event recognition can be regarded as a category of scene classification. SUN database [21] and MIT67 dataset [13] are examples of scene classification datasets which cover various scenes of different categories. There are mainly two approaches for scene classification. The first approach is to exploit Bag-of-Features(BoF) representation [2]. The classification performance can be improved when BoF model is combined with spatial pyramid matching [11] or its variants [22]. The second approach is to classify scenes through CNN. Donahue et al. [4] extracted features from pre-trained CNN and used the features for the scene classification. In [9], spatio-temporal information is learned by extending a CNN in time domain to classify videos. This method is also applied to human action recognition.

While human action recognition or pose estimation has drew lots of attention [3, 5, 19], recognizing cultural event from a single image has been rarely studied. In addition, while the previous scene classification datasets contain images of wide area, some of the images in the cultural event classification dataset contain a closeup of people or specific objects. Each cultural event has its own distinctive characteristics, which can be garments, pose of human, event-specific objects such as beer glasses, or other semantic features. Rather than extracting those information explicitly, we trained CNN with training examples to learn the discriminant features automatically.

3. CNN with region proposals

In the images of cultural events, there are regions that distinguish one event from the others. In other words, visual characteristics of a cultural event appear only on the subregions of the image. For instance, images of *Oktoberfest* contain beer glasses, and images of *Sendfest* contain lots of sand textures. The motivation of our method is that training and testing with only the discriminant regions will improve the accuracy of the classification. However, it is hard to locate the regions that contain key information of the image. Inspired by [6], we extract region proposals which are candidates of the distinctive regions for cultural event recognition. We will refer to the extracted region proposals as *image regions* in the paper. The image regions are subimage of the original image whose size and location are defined by a rectangular box. The details of our algorithm is explained in the following subsections.

3.1. Extracting region proposals

To extract distinctive and meaningful regions from an image, the extracted regions should be repetitively detected and need to be robust to scale or rotation variations. Though we have no idea of which features are discriminant for cultural event recognition, we expect that the discriminant features are in the object levels. For these reasons, we used [20] to obtain the region proposals which is also used in [6] for object detection tasks. As in [20], possible object locations are extracted via selective search which combines an exhaustive search and segmentation. From the extracted candidate regions, we exclude the ones that is too small (regions whose width or height is less than 10% of the original image) and too tall or fat regions (regions whose width/height ratio is grater than 2 or less than 0.5). After the exclusion step, approximately 200 to 300 image regions are extracted from one image. Data augmentation is another advantage of region proposal extraction. It is able to generate over one million training examples from the thousands of training images, which is sufficient to train our deep CNN.

3.2. Structure of CNN

Using the patches extracted in Section 3.1, CNN is trained to classify each patch. The structure of our CNN is illustrated in Figure 2. As shown in the figure, our model consists of 3 convolutional layers, each of which is followed by the corresponding pooling layers (3 pooling layers), and 2 fully connected layers. The filter size of each convolutional layer is 5×5, 3×3, and 3×3 respectively, and the number of filters for each convolutional layer is 96, 128, and 128 respectively. First convolutional layer is applied on the input image with stride of 2, and images in the second and third layers are padded with 1 on both sides before applying convolution. All pooling layers pool over 3×3 regions with stride of 2. Hence, there are overlapping regions

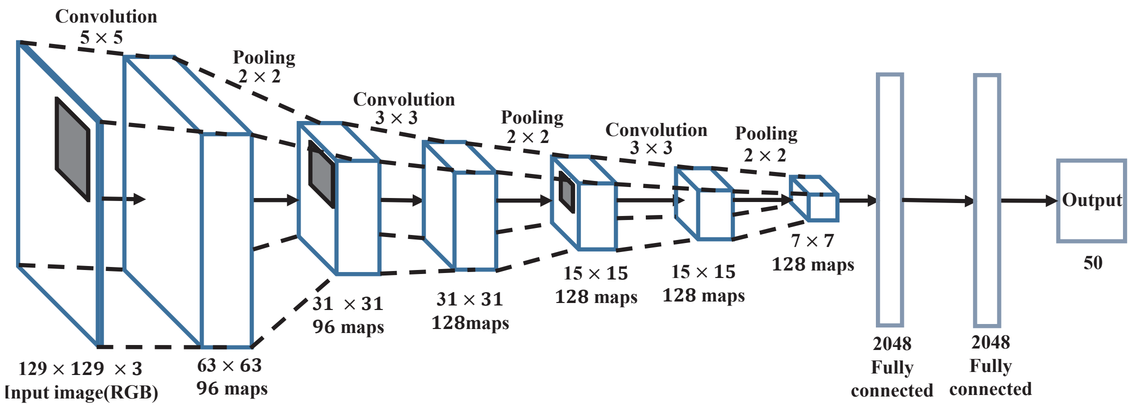


Figure 2: Structures of CNN. Filters at the first convolutional layer is applied with stride of 2, and images at the second and third layers are padded with 1 pixel. Dropout is applied to fully connected layers.

for each pooling kernel. There are two fully connected layers, each of which has 2,048 nodes. In the fully connected layers, dropout [17] is applied to improve generalization performance. Rectified Linear Unit(ReLU) is used as the activation function of the network. Finally, softmax classifier is applied on the last layer, where the cross entropy is used as a loss of the model [12]. Caffe framework [8] is used to implement, train, and test our model. Each image region is resized to 133 × 133, and the resized region is randomly cropped to 129 × 129 size. In practice, random cropping during the training prevents overfitting and allows robustness on the small translation [10, 15]. The image region is cropped at the center during the testing. The cropped images is fed to the CNN as an input after mean subtraction. Considering number of images in the training set, number of filters for each layer is adjusted to approximately half the size of the model of [10]. Recently, CNN with small filter size showed promising results [16], so we used relatively small filter size in the convolutional layers. The proposed CNN is able to classify the training examples efficiently without pre-training step.

3.3. Training and testing CNN

The Chalearn cultural recognition database has 5,875 training, 2,332 validation, and 3,569 test images, respectively. During the validation phase, only the training set is used for training CNN, while both training and validation sets are used to train CNN in the test phase. The number of regions extracted is 1,558,815 for train sets, 624,629 for validation sets, and 951,738 for test sets, respectively. During the training, the cost function is minimized by stochastic gradient descent method with learning rate of 0.01, momentum of 0.9, and weight-decay of 0.0005. The minimization process is performed for 250,000 iterations with mini batch size of 128, which corresponds to around 20 epochs when using only the training set and around 15 epochs when using both the training and validation sets. It took 16 to 17 hours to train the CNN on a single desktop PC with 24GB RAM and GeForce Titan Black GPU.

Testing is applied for every extracted region through the trained CNN. Therefore, class probabilities are assigned for each region, and the assigned class probabilities of regions from the same image is gathered together to produce the final prediction of the image. The method how to predict the final probabilities of an image is explained in the next section.

4. Classification of an image

4.1. Rejecting high entropy regions

Before gathering the classification results of each region, the image regions whose class probability distribution have high entropy are discarded. The entropy of probability distribution $X = [x_1, \cdots, x_C]$ is calculated as

$$\mathrm{H}(X) = -\sum_{i=1}^{C} \mathrm{P}(x_i) \ln \mathrm{P}(x_i), \tag{1}$$

where C is the number of classes and x_i is the probability that the region belongs to class i. Note that the maximum entropy for the cultural event recognition database is about 3.9 since there are 50 classes. The threshold is determined as 2.5 by testing on the validation set, which is fixed for the whole experiments in this paper. The regions that have high entropy values are considered to contain not much discriminant information, and they are not used for classification of the input image.

4.2. Combining the classification probabilities of image regions

Several simple schemes are tested for the classification of an image from the class probabilities of each region of the image.

Mean of probabilities: As the simplest approach, the class probabilities of an input image can be calculated as the average class probabilities of all the subregions. Hence, the probabilities of each class for the input image $\mathbf{x}$ is assigned as

$$\mathrm{P}(\mathrm{class} = c|\mathbf{x}) = \frac{\sum_{i=1}^{N} \mathrm{P}(\mathrm{class} = c|r_i)}{N}, \qquad (2)$$

where r_i is an image region ($i = 1, 2, \ldots, N$), N is the number of image regions from the input image $\mathbf{x}$, and c is the class for cultural events.

Weighted sum of probabilities: This scheme weights the image regions that cover large areas of the input image. The underlying intuition for this scheme is that large regions may contain more information than smaller regions. The area, or the number of pixels, of the image region is used as the weight. Then, the calculated weighted sum is normalized as in the following equation,

$$\mathrm{P}(\mathrm{class} = c|\mathbf{x}) = \frac{\sum_{i=1}^{N} \mathrm{area}(r_i)\mathrm{P}(\mathrm{class} = c|r_i)}{\sum_{c=1}^{C}\sum_{i=1}^{N} \mathrm{area}(r_i)\mathrm{P}(\mathrm{class} = c|r_i)}, \qquad (3)$$

where $\mathrm{area}(r_i)$ denotes the area of the image region r_i.

Maximum probabilities count: This method counts the label that has maximum probability for each image region. The probability for each class is calculated as

$$\mathrm{P}(\mathrm{class} = c|\mathbf{x}) = \frac{\sum_{i=1}^{N} \mathrm{I}(\operatorname*{argmax}_k \mathrm{P}(\mathrm{class} = k|r_i) = c)}{N}, \qquad (4)$$

where $\mathrm{I}(\cdot)$ is the indicator function which takes on zero (miss) or one (hit). Hence, this method ignores the probability distribution of each region, and only the label that has maximum probability contributes to the classification.

5. Experimental results

In the Chalearn cultural event recognition challenge, performance is measured in terms of mean average precision (mAP). mAP is calculated as the average area under the precision-recall curve for each class. The mAP of our method on validation set using various prediction schemes in Section 4.2 is shown in Table 1. The three methods are abbreviated as mean, weighted sum, and counting respectively. Whether the entropy thresholding (ET) is applied (w/ ET) or not (w/o ET) is also indicated. All methods outperform the baseline method and show similar performance. The simplest approach, mean of the probabilities, shows best result among the three. The performance of mean of probabilities with ET is slightly better than without it, but the difference is marginal.

Method	mAP
Mean w/ ET	**0.683**
Mean w/o ET	0.677
Weighted sum w/ ET	0.671
Counting w/ ET	0.661
Baseline method [1]	0.239

Table 1: Mean average precision on the validation set

Method	Top-1 acc.	Top-5 acc.
Image region w/o ET	43.0%	68.2%
Image region w/ ET	50.7%	75.1%
Mean w/ ET	**69.9%**	87.7%
Weighted sum w/ ET	69.1%	**87.8%**
Counting w/ ET	69.8%	87.2%

Table 2: Classification accuracy on the validation set

We also measured the classification accuracy of our methods. Both top-1 accuracy and top-5 accuracy are reported in Table 2. In the table, the first two rows show the average classification accuracy of total 624,629 image regions for the validation set. This accuracy is directly related to the performance of the trained CNN. The image regions with high entropy have little information on the specific cultural event and we can see that the image region classification performance is increased by around 7% by excluding high entropy regions in both top-1 and top-5 accuracies. The lower 3 rows in Table 2 show the image classification performances by combining the class probabilities of each image region as described in Section 4.2. All combining schemes show nearly the same classification accuracy: 69% for the top-1 accuracy and 87% for the top-5 accuracy. Interestingly, image classification accuracy is boosted significantly compared to the image region classification result. Top-1 accuracy is increased by 20% and top-5 accuracy is increased by 12%. Therefore, it can be said that, by classifying multiple subregions from the image and aggregating the results together, the classification accuracy can be boosted.

Using the trained CNN, one can find examples of image regions that have high class probability for one class. Figure 3 shows the image regions that have class probability of larger than 0.99 for a specific cultural event. This means that the images in Figure 3 is typical examples of regions that contain distinctive information from the other classes. It can be seen that some images cover wide scenes while others contain closeup of specific objects. Therefore, it can be argued that region extraction in various size influence the performance. Also, by scrutinizing the regions with high probability (or low entropy), the visual characteristics of each cultural event can be easily identified.

Team name	mAP
MMLAB	0.855
UPC-STP	0.767
MIPAL_SNU	**0.735**
SBU_CS	0.610
MasterBlaster	0.582
Nyx	0.319

Table 3: Mean average precision on the test set

Lastly, Table 3 shows the mean average precision on the test set, which is the final result of the cultural event recognition challenge. Image regions from both the training and the validation sets are used as training examples. The method used for generating the result of the test set is mean of probabilities with entropy thresholding. The increased performance compared to the result on the validation set can mainly be attributed to the increased number of training image regions. All participants are denoted as their team names, and our method, MIPAL_SNU, ranked 3rd among 6 participants. Though our method is inferior to the best method, it has some desirable aspect that it does not require prior knowledge or pre-training.

6. Conclusion

In this paper, a cultural event recognition algorithm is proposed based on the extraction of region proposal and convolutional neural networks. The proposed CNN successfully classified training images which are the subregions of images. The classification probabilities of image regions in an image are combined to generate the final image classification result. Our framework is simple, yet powerful enough to achieve the competitive result compared to the other methods of the challenge. Also, due to the generality of the proposed algorithm, it can be applied to any other image classification problems as well as the cultural event recognition.

Though simple averaging or counting scheme reasonably improved the classification performance, developing more efficient combining scheme should be considered as a future work. Also, filtering out unnecessary image regions during the training can improve the classification accuracy of the CNN.

References

[1] X. Baro, J. González, J. Fabian, M. A. Bautista, M. Oliu, I. Guyon, H. J. Escalante, and S. Escalers. Chalearn looking at people 2015 cvpr challenges and results: action spotting and cultural event recognition. In *CVPR, ChaLearn Looking at People workshop*, 2015. 1, 4

[2] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *Workshop on statistical learning in computer vision, ECCV*, volume 1, pages 1–2. Prague, 2004. 2

[3] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. *arXiv preprint arXiv:1411.4389*, 2014. 2

[4] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. *arXiv preprint arXiv:1310.1531*, 2013. 2

[5] S. Escalera, X. Bar, J. Gonzlez, M. A. Bautista, M. Madadi, M. Reyes, V. Ponce, H. J. Escalante, J. Shotton, and I. Guyon. Chalearn looking at people challenge 2014: Dataset and results. In *ECCV Workshops*, 2014. 2

[6] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 580–587. IEEE, 2014. 2

[7] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. *arXiv preprint arXiv:1502.01852*, 2015. 2

[8] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014. 3

[9] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 1725–1732. IEEE, 2014. 2

[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 2, 3

[11] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, pages 2169–2178. IEEE, 2006. 2

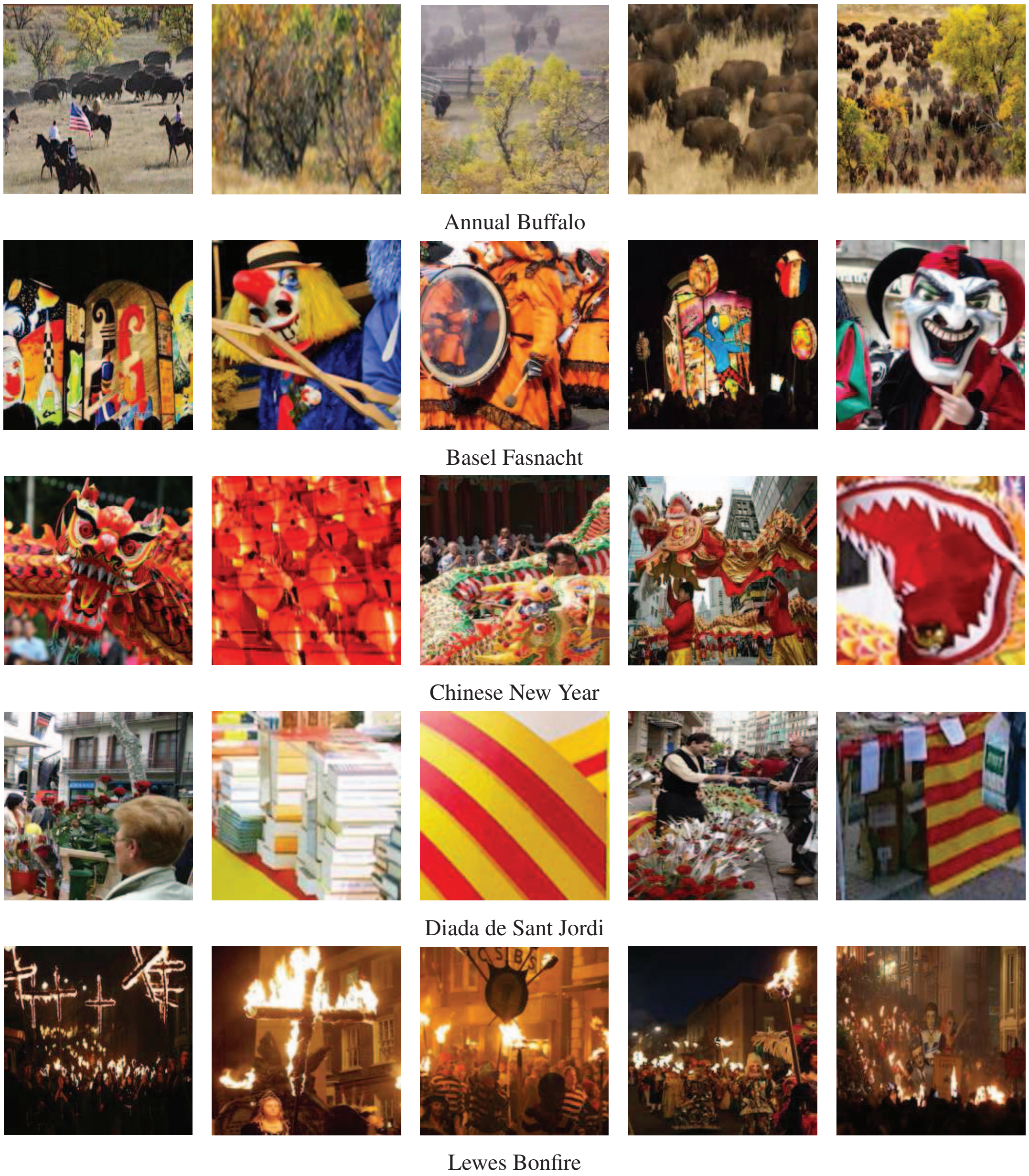
[12] K. P. Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012. 3

[13] A. Quattoni and A. Torralba. Recognizing indoor scenes. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 413–420, June 2009. 2

[14] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge, 2014. 1

[15] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. *arXiv preprint arXiv:1312.6229*, 2013. 3

[16] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 3



Annual Buffalo

Basel Fasnacht

Chinese New Year

Diada de Sant Jordi

Lewes Bonfire

Figure 3: The examples of image regions that have class probability of larger than 0.99 for each cultural event.

[17] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014. 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. *arXiv preprint arXiv:1409.4842*, 2014. 2

[19] A. Toshev and C. Szegedy. Deeppose: Human pose estimation via deep neural networks. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 1653–1660. IEEE, 2014. 2

[20] J. R. Uijlings, K. E. van de Sande, T. Gevers, and A. W. Smeulders. Selective search for object recognition. *International journal of computer vision*, 104(2):154–171, 2013. 2

[21] J. Xiao, K. Ehinger, J. Hays, A. Torralba, and A. Oliva. Sun database: Exploring a large collection of scene categories. *International Journal of Computer Vision*, pages 1–20, 2014. 2

[22] L. Xie, J. Wang, B. Guo, B. Zhang, and Q. Tian. Orientational pyramid matching for recognizing indoor scenes. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 3734–3741. IEEE, 2014. 2