Urban Tribe

Anonymous WACV submission

Paper ID ****

Abstract

Recognition of social styles of people are an interesting but not yet popular topic. In this paper, we explore the use of pre-trained convolutional neural network(CNN) features for social group recognition. A social group recognition framework is proposed. Pre-trained CNN is fine-tuned, and CNN features are used for both individual person images and global scene images. Our model shows promising results on urban tribes dataset, with 71.23% accuracy, which significantly outperforms previous results of 46%. We also find that there is a correlation between the probability of an image being in Imagenet classes and social group classes, and that better-recognized categories have more highly-correlated ImageNet categories. This gives us insight into the features extracted from pre-trained CNNs.

1. Introduction

In the past few years, there have been impressive progress in understanding semantic meaning of images, such as object recognition, scene recognition, and object detection. The power of Convolutional Neural Networks (CNNs) has is especially noticeable. However, the analysis the social features of images of groups of people has not attracted a great deal of research. Current search algorithms fail to capture information of personal styles or social characteristics of groups of people, but retrieve images with similar global appearance [1]. The analysis of groups of people is difficult in that the group categories are semantically ambiguous, and have high intra-class variance.

Recognition of groups of people from a social perspective provides many potential applications. With more accurate group searching results, more accurate recommendations can be made in social networks, and more relevant advertisement for particular groups of people benefits both consumers and sellers.

Kwak *et al*. studied this problem of group recognition ([1], [2]). They created an urban tribe dataset consisting of 11 categories, with about 100 labeled per class. They proposed a group recognition pipeline. Rather than classifying isolated individuals in the group images, they focused on group features and models.

CNN architecture has been proved to achieve outstanding results in various computer vision tasks, and it's argued that deep architectures in CNN can capture visual features of different semantic level in the hidden units ([3]). Recently, features learnt for large scale recognition tasks with large amounts of training data have been used for new tasks, and the features outperform many conventional features in many tasks([3], [4]). This indicates that generic visual features may be obtained from pre-trained CNN model.

not sure about this paragraph The most frequently used pre-trained model was proposed in [5] trained with ImageNet dataset[1]. And the new tasks for which CNN features are used range from object recognition, scene recognition, to subcategory recognition. However, the recognition of style hasn't been researched yet. The social groups of individuals require the recognition of people's social characteristics, which is a problem of style, instead of category. Moreover, the ImageNet classes used for pre-training cover few human images, making the problem more interesting.

In this paper, we investigate the generalization ability of pre-trained CNN features to social group recognition. We propose a CNN feature based architecture for social group recognition. Our model takes in both individual features and global scene features fine-tuned from CNN pre-trained weights. Our result shows a boost of performance from the previous classification method provided by [1]. We show that both individual information and global scene information contribute to a social group's characteristics, and that different feature extraction schemes for individual and global information is necessary. We also show the role of adapting generic pre-trained CNN features to social group images with fine-tuning.

We further investigate why features extracted from pre-trained CNN are useful for the urban tribe recognition task. For an input image, there is a correlation between the probability of it being in ImageNet classes and being in urban tribes classes. Moreover, the degree of correlation is related to the recognition rate of different urban tribes classes

[1] http://image-net.org/challenges/LSVRC/2012/browse-synsets

- better-recognized categories have more highly-correlated ImageNet categories. This may indicate that the "generic" features extracted by pre-trained CNN networks are not so generic. The network is trained to separate the ImageNet classes most, and if we use the features for a new classification task, the performance of the task is related to how well the new classes can be "mapped" to the ImageNet classes. However, the actual relationship between the two types of categories is still mysterious in most cases.

2. Related Work

Convolutional Neural Networks(CNNs) with back-propogation were introduced around 1990's by LeCun *et al*. [6]. Since then, CNNs have shown successful results on various computer vision tasks, such as hand-written digit classification ([7],[8]), ImageNet Large Scale Visual Recognition Challenge ([5], [9]), object detection ([9], [10], [11]), object localization ([10]), and many modified architecture of CNNs have been proposed([8], [12], [13]).

Recently, many researchers have shown the utility of generalization of pre-trained CNN features on large dataset such as ImageNet. Krizhevsky *et al*. showed excellent generalization of the pre-trained CNN features[3]. They kept all the layers of ImageNet-trained model using fixed pre-trained features except for the last softmax classifier, and achieved best results on Caltech-101 and Caltech-256. Donahue *et al*. used different layers of pre-trained CNN network as features and trained simple classifiers such as SVM and Logistic Regression, and outperformed the state-of the-art on several vision challenges such as scene recognition and domain adaptation[4] . Karayev *et al*. compared different approaches for photographic style recognition, and pre-trained CNN features generally obtain best result [14].

Kwak *et al*. created an urban tribe dataset consisting of 11 classes[1]. The classes are defined from social group labels provided by Wikipedia. They selected the eight most popular categories from their list of subculture, and added three other classes corresponding to typical social venues in addition. For each class, images of groups of people were discovered with different search engines, and a broad range of images for each class were collected. Kwak *et al*. also provided a group description and several classification methods([1], [2]). Group description consists of person descriptors and global group descriptors: Six part of person is detected, and a set of predefined descriptors are computed for each part, including ratio of skin pixels, color information like RGB histograms, and HoG features; Global descriptors use a both low level and high level descriptors to describe the context and group properties of the image. Low lever features include color information, Gist, HoG and ratio of pixels of person, and high level descriptors includs proximity of persons, alignment or pose of the group, and scene layout of individuals. Two options of classification methods are provided: bag of parts-based classification and SVM-based classification.

Categorizing the social groups of individuals belongs to fine-grained classification task which recently draws more interest of the computer vision community. It aims at giving the fine-grained categories in a certain class. Fine-grained classification is more difficult than conventional classification tasks, because the categories are semantically as well as visually similar, and are even challenging for humans. The Fine-grained Challenge 2013 (FGComp) provided the data in several categories including aircraft, birds, cars, dogs, and shoes. [15] achieved the best result using classifier based on fisher vectors. However, CNN based methods using [16] or [4] gave inferior results, especially when the bounding box of test data is unknown.

There is some research in analyzing social groups of people. [17] showed the visual structure of a group helps understanding events. [18] showed social relationships modeling helps people recognition. [19] used both local and global factors for group level expression analysis.

3. Methods

This section describes the urban tribe dataset and elaborates on the model architecture.

3.1. Urban tribes dataset

Urban tribes are groups of people who have similar visual appearances, personal style and ideals. The urban tribes dataset consists of 11 different categories: *biker, country, goth, heavy-metal, hip-hop, hipster, raver, surfer, club, formal, casual/pub*, with an average of 105 images from each category.

Unlike conventional visual classification problems, urban tribe categories are more ambiguous and subjective. Also, each class contains a broad range of scenarios. The high intra-class variation of the urban tribe dataset makes the classification task challenging. The urban tribe dataset also has some interesting properties. The number of people in each urban tribe image varies. Members in one tribe often have similar visual styles, including their clothes, accessories, and even demeanor. For example, surfers possibly carry surfboards, and the goth often have dark attire, makeup and hair. The environment they are in also contributes to each tribe characteristics: pictures of country tribes are more likely to be taken outdoors with grassland, while pictures of clubbers are often photographed in clubs with dim lightings.

3.2. Classification hierarchy

To utilize the properties of urban tribes fully, our feature vector consists of both elements: individual features and environmental features. For each feature type, we use

a similar extraction strategy. Individual features and environmental features are hierarchically combined to form the final decision function. The network hierarchy is shown in Figure 1.

For each group image, we represent the group G as the combination of a set of people and the environment of the scene. To give the prediction of class C, the individual feature vectors and scene feature vectors are extracted separately.

For the individual feature vectors, first, individual candidate person images are detected with a poselet based person detection algorithm. The candidate person images $H = \{H_1, H_2, ..., H_p\}$ are used as a whole instead of a set of body part bounding boxes. Each candidate person is resized to 256×256, and ten 227×227 patches $\{h_{ij}\}, i \in \{1, 2, .., p\}, j = 1, 2, ..., 10$ are extracted (patches from four corner and the center, and their horizontal reflections).

Each Individual image patch h_{ij} then passes through the Convolutional Neural Network for person images CNN_{Person}, generating activations from the 6th and 7th hidden layer. The activations from 6th and 7th layer are both 4096 dimensioned. They are concatenated to form an 8192-dimensional vector f_{ij}, where $i \in \{1, 2, .., p\}, j \in \{1, 2, ..., 10\}$.

The feature vectors are then fed into a multi-class SVM_{Person}. We use LIBLINEAR[20] to train the SVM on individual patches, and to estimate probabilities for each category given individual patch h_{ij}: $\Pr_{ij}(C|h_{ij}), C \in \{1, 2, .., c\}$, where c is the number of classes in urban tribe dataset. The individual patches h_{ij} in one group image are usually highly correlated. Therefore, in order to obtain a reliable probability estimate from the noisy yet correlated set of probabilities $\Pr_{ij}$, a simple but effective average pooling is performed to $\Pr_{ij}$:

$$\Pr_{People}(C|H_1, ..., H_p) = \frac{1}{10p} \sum_{i,j} \Pr_{i,j}(C|h_{ij}) \quad (1)$$

$\Pr_{People}(C|H_1, ..., H_p)$ is the probability estimate of class C given the set of people candidate images H.

On the other hand, the entire environment in the scene image, denoted by S, is directly utilized for probability estimation. The procedure to generate probability estimate of class C given the environment as a whole $\Pr_{Scene}(C|S)$ is similar with that for $\Pr_{People}(C|H)$. The difference is that the input of Convolutional Network is 227×227 patches extracted from the entire scene image, and the fine-tuned Convolutional network: CNN_{Scene} and SVM: SVM_{Scene} are trained with the training set of entire scene images. Several different strategies to extract patches from scene images and corresponding Convolutional Neural Network architectures are explained in Section 3.3.

Therefore, the probability estimate of a class C given observation of scene S is:

$$\Pr_{Scene}(C|S) = \frac{1}{K} \sum_{k=1}^{K} \Pr_k(C|s_k) \quad (2)$$

where K is the number of scene patches extracted from one group image, and this number varies with different patch extraction strategies. s_k is the kth scene patches. $\Pr_k(C|s_k)$ is the probability for class C given kth scene patch. Average pooling is still used here, because the assumption of high correlation in patches holds.

Now we have the estimates of two kinds of conditional probability $\Pr_{People}(C|H)$ and $\Pr_{Scene}(C|S)$. We make a strong assumption that the two types of features are independent, and that the prior probability distribution of the urban tribes $\Pr(C)$ is a uniform distribution. The classification problem can be expressed as maximizing the objective function:

$$L = \arg\max_{i=1,...,c} \Pr(C = i|G) \quad (3)$$

where

$$\begin{aligned} \Pr(C = i|G) &= \Pr(C = i|H, S) \\ &= \frac{\Pr_{People}(C = i|H_1, ..., H_p) \cdot \Pr_{Scene}(C = i|S)}{\Pr(C = i)} \\ &\propto \Pr_{People}(C = i|H_1, ..., H_p) \cdot \Pr_{Scene}(C = i|S) \end{aligned} \quad (4)$$

and L is the predicted label for the group image.

3.3. Convolutional network feature extraction

It is shown in many experiments that a set of weights of convolutional network trained from ImageNet can generate a set of generic visual features.

Following [16]'s work, we use the network framework called Caffe. The network architecture is described in [5], which won the ImageNet Large Scale Visual Recognition Challenge 2012. We take the activations from the 6th and 7th hidden layer of the convolutional neural network, which are two fully connected layers before the class prediction layer. We also take the activations from 6th or 7th layer alone as comparison. We choose these two layers, because as the layers ascend, the features extracted show increasing invariance and semantic meaning.

We use the pre-trained set of weights of the network released by Caffe as the initial parameters of our network. The pre-trained model was trained on ImageNet ILSVRC-2012, and all images are first resized to 256×256 before they can be used as inputs to the network.

3.3.1 Pre-processing of the dataset

The urban tribe dataset is a relatively small dataset, and both people candidate crops and scene images are of various res-

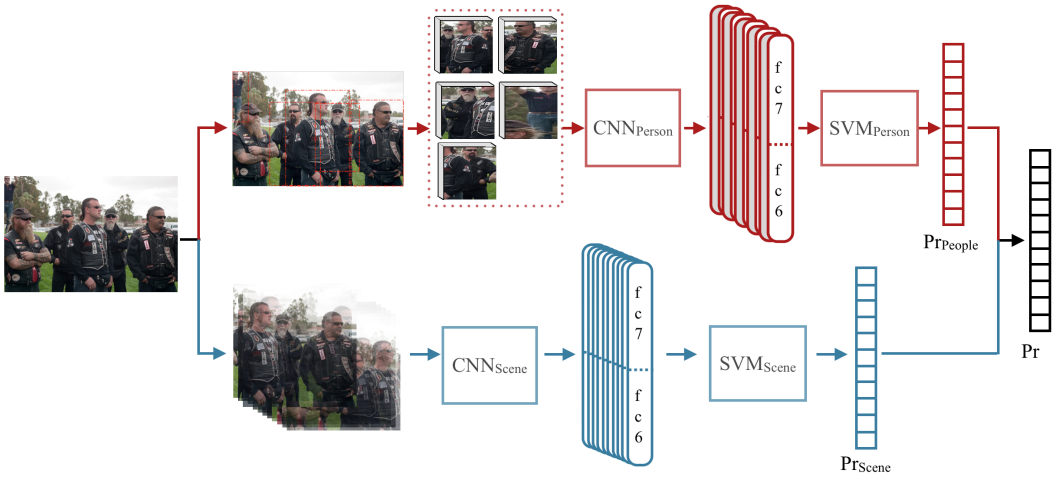


Figure 1: Architecture of classification algorithm using Nets$_{SDense}$ which is introduced in Section 3.3.3. The upper half estimates the probability given people candidate images, and the lower half estimates the probability given the entire scene. Dense crop CNN$_{Scene}$ and distorted crop CNN$_{Person}$ are used, as shonw in Section 3.3.1.

olution. Our convolutional neural networks requires constant input image size of 227 × 227, so pre-processing of the dataset is necessary.

There are several strategies to make one image compatible with the CNN:

1. *Distort cropping:* As in [5], resize the image to a fixed resolution of 256×256, and crop five 227×227 patches (from four corner and the center) and their horizontal reflections to generate ten patches from one single image. This way, the aspect ratio of the original images are lost, but for each crop, the portion it takes from the original image is fixed, so that the amount of information all the crops have is relatively stable.

2. *Sparse cropping:* Keep the aspect ratio of the original image, resize the shorter side to 256, and then crop five 227 × 227 patches and their horizontal reflections as mentioned. This method avoids distortion of the image and objects in it, but the crops will possibly lose much information when the aspect ratio of the original image is far away from 1.

3. *Dense cropping:* Keep the aspect ratio of the original image, resize the shorter side to 256, and then densely crop multiple 227 × 227 patches and their horizontal reflections. This way, the information of original image is kept by dense cropping process, and the distortion is avoided. The number of crops attained with this method is larger than the previous two methods.

3.3.2 Network Fine-tune

Although the pre-trained network from [16] can already generalize well to many datasets, the urban tribe dataset has its unique property. It emphasizes certain visual features such as certain clothing styles, while pays less attention to other visual features. Also, it emphasizes style of the object rather than distinct category. To rearrange the importance of different features and adjust the features to adapt to urban tribe dataset, the network can be fine-tuned.

The dataset used for fine-tune is the same set used for SVM training. The input patches of the Convolutional network is of size 227 × 227. The initial Convolutional network has 1000 outputs in final layer, corresponding to 1000 class-wise probability predictions. In our fine-tune process, the last layer is replaced by 11 probability prediction outputs, and the initial weights of the last layer connection are initialized to have zero mean gaussian distribution. Back propagation is used, and the learning rate is set to be small so that the fine-tune process adapts the extracted features to urban tribe dataset while preserving the initial property in general: the initial learning rate used for pre-training is 0.01; We set the initial fine-tune learning rate of the parameters except for the last layer as 0.001, and keep the initial learning rate for last layer 0.01, because the last layer is not pre-trained.

3.3.3 Choices of network combination

Scene images and individual images have different properties, and need different strategies of pre-processing and sep-

arate fine-tuning. For scene network CNN_{Scene} and scene images, due to the small size of the dataset, we use the *dense cropping* technique, the third technique in Section 3.3.1, to increase the dataset. For person network CNN_{Person} and corresponding input, we use the *distort cropping* technique, because the subimages have normally long height and short width, and the second and third strategies using squared crops of a person image will lose much information, no matter which location we choose to crop them; whereas the first method ensure each crop keeps the essential features for classification.

The combination of dense crop CNN_{Scene} and distorted crop CNN_{Person} are denoted as Nets_{SDense}.

We also construct other combination of networks for comparison:

1. Nets_{NoTune}: Directly use the pre-trained network by [16] for both scenes and persons, and use the *distort cropping* technique (distorted crops) as input patches for both networks. This choice of cropping strategy is in consistent with the way the network is pre-trained.

2. $\text{Nets}_{SSparse}$: Use the *sparse cropping* strategy for scene features, and the *distort cropping* strategy for person features. .

3. $\text{Nets}_{SDistort}$: Use the *distort cropping* strategy for both scene features and person features.

4. Experiments and Results

In this section, the performance of the proposed classification scheme is evaluated and analyzed. In the experiments, six rounds of 5-fold cross validation are performed, therefore we have 30 training experiments in total. Dataset is partitioned into 5 equal sized subsets, containing one-fifth of the data points from each category. One of the subsets is used as test set, and the remained 4 subsets are used as training data. In the fine-tune procedure, there are 6000 training iterations, and learning rate is decreased by ten times after each 1000 iterations.

4.1. Urban tribe classification performance

Table 1 shows the comparison of performance using different approaches. The 30 segmentations of datasets are used for all the approaches tested in this section, and 30 test results are averaged for each approach. The standard error is shown with the accuracy in Table 1. We also compare our result with the result achieved by [1] using their best model. The advantage of CNN pre-trained features is obvious.

The confusion matrix is shown in Figure 2 for Nets_{SDense}, and all the 30 training experiments are averaged. We can observe there is a obvious difference of difficulty of different categories. Class *formal* has accuracy as high as about 90%, while class hipster is the most difficult class, having less than 60% accuracy.

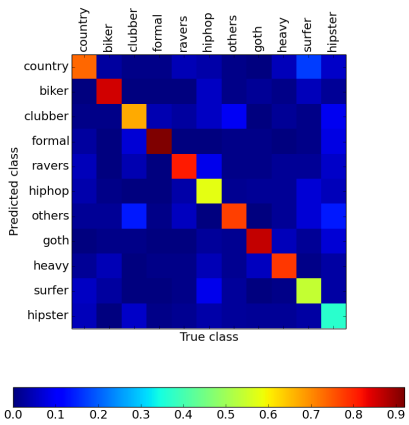


Figure 2: Confusion matrix for classification results with Nets_{SDense}, using people and scene features.

Comparing the result of using different features in the same approach shows the necessity of every step of our architecture. In results using Nets_{SDense} with concatenated features, average accuracy for each candidate person is low as 47.10%. Average pooling of candidate person probability estimates produces a large accuracy increase of about 20%. Accuracy using the entire scene only results in 67.26% accuracy. Combining probabilities $\text{Pr}_{People}(C|H_1, ..., H_p)$ and $\text{Pr}_{Scene}(C|S)$ achieves accuracy as high as 71.22%, which verifies the complementary role of people candidate feature and environment feature in a group image.

We also compare vertically the accuracy of different approaches, to show the role of network feature concatenation. Using only 7th layer or 6th layer activation from the networks Nets_{SDense} produces decent results, showing both layers' activations can generate high semantic features. Concatenating both layers' activations increases the accuracy by 0.5%, indicating the slight information loss of the 7th fully connected layer.

To see the role of fine-tuning, we can compare the result of Nets_{NoTune} and $\text{Nets}_{SDistort}$. These two approaches both use resizing that causes distortion, and they only vary in fine-tune procedure. There is a large performance improvement with fine-tune, both for person features and scene features. This shows the benefit of adapting existing generic Convolutional Network to specific dataset.

$\text{Nets}_{SDistort}$, $\text{Nets}_{SSparse}$, and Nets_{SDense} use different patch extraction strategies. Note that we use the same distorted patch extraction method for person images, as mentioned in Section 3.3.3, while we use three different meth-

Table 1: Performance of different approaches using different information.

Accuracy (%)	Individual candidate	People	Entire scene	People+Scene
Nets$_{NoTune}$ with concatenated features	40.03 ± 0.31	64.18 ± 0.63	62.61 ± 0.58	69.19 ± 0.51
Nets$_{SDense}$ with fc7 features	46.95 ± 0.33	67.42 ± 0.50	64.87 ± 0.39	70.68 ± 0.47
Nets$_{SDense}$ with fc6 features	45.80 ± 0.37	66.31 ± 0.47	66.72 ± 0.56	70.68 ± 0.44
Nets$_{SDense}$ with concatenated features	$\mathbf{47.06 \pm 0.37}$	$\mathbf{67.46 \pm 0.51}$	$\mathbf{67.01 \pm 0.52}$	$\mathbf{71.45 \pm 0.48}$
Nets$_{SSparse}$ with concatenated features	$\mathbf{47.06 \pm 0.37}$	$\mathbf{66.48 \pm 0.41}$	66.81 ± 0.42	71.26 ± 0.49
Nets$_{SDistort}$ with concatenated features	$\mathbf{47.06 \pm 0.37}$	$\mathbf{67.46 \pm 0.51}$	65.06 ± 0.48	71.15 ± 0.53
SVM_8[1]	-	-	-	46(std: 2)

ods for scene images. The results for scene images show the advantage of keeping the aspect ratio of scene images, and the slight advantage of using dense crops. However, the final results with People+Scene for the three methods don't have significant differences, this is due to the combination with people information.

4.2. Convolutional Network feature analysis

Yufei: Is this section necessary? We use t-distributed stochastic neighbor embedding technique [21] to visualize the power of different layers' features in convolutional network. Two dimensional embedding of the high dimensional CNN feature space is extracted, and we plot the data as 2-d points with different colors indicating different classes they belong to. Powerful features will pull data of different semantic classes more apart.

We randomly choose one training-test partition and corresponding fine-tuned network parameters, and to avoid overfitting effect, we examine the test set of Nets$_{SDense}$ approach. In Figure 3, each data is plotted as a dot in each figure. The three columns correspond to data separation in first layer, fourth layer and seventh layer respectively. The first row visualizes features of all test data in scene CNN CNN$_{Scene}$, and second row picks three classes from the first row. The third row is features of CNN$_{Person}$, and the three classes are picked for the last row. The three classes chosen in second and last row is (goth:red, heavy:green, surfer:blue).

In both CNN$_{Scene}$ and CNN$_{Person}$, there is a clear trend of class separation. As the layer ascends, , the data from same class are more concentrate, and inter-class distance are larger. The selected three classes show the trend more clearly: they essentially form three clusters in the seventh layer. As shown in the second and last rows, person data is more challenging: The semantic classes are not as separate as with the scene features.

4.3. Urban tribe classes vs. ImageNet classes

It's recently being acknowledged that CNN pre-trained features are generic and can be used for new tasks. In this section we check the relationship between the new tasks and

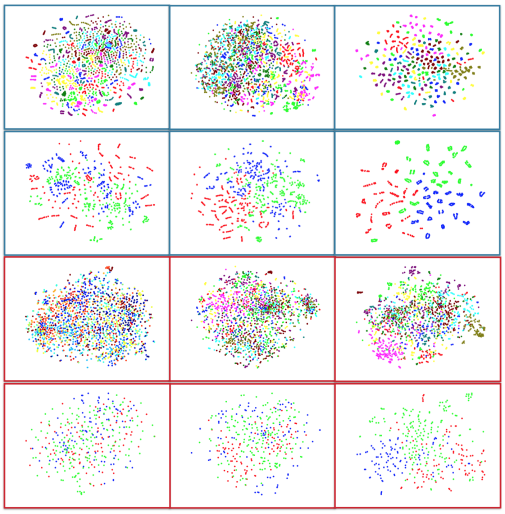


Figure 3: Feature visualization of Nets$_{SDense}$: CNN$_{Scene}$ (first two rows) and CNN$_{Person}$(second two rows). Rows: first: all scene images; second: all data from goth/heavy/surfer urban tribes; third: all test person images; fourth: all test data from goth/heavy/surfer urban tribes. Columns: first: first convolutional layer; second: fourth convolutional layer; third: seventh fully convolutional layer

original ImageNet task, using the urban tribe dataset,which gives us some insight about the features extracted from pre-trained network.

The urban tribe dataset contains groups of people, and the important features for categorization are mainly human related features, such as attire, make up, posture and expressions. However, the ILSVRC dataset used for pre-training contains few images of humans. Instead of examining the output of the layers directly, we try to find the relationship between the 1000 classes in ILSVRC dataset and the classes in two image sets: urban tribe dataset, and candidate person images extracted from urban tribe dataset.

We use the parameters of pre-trained CNN network as our feature extraction model, and train a softmax layer on top of its 7th layer to predict the probabilities of one input

image(either scene image or candidate person image) being in certain urban-tribe class $\Pr(l_{urban})$, where l_{urban}pre-trained is the 11 urban tribe categories. The output layer is trained for 3000 training iterations. We can also use the output of the pre-trained CNN network to predict probabilities of one input image being in certain ImageNet category, denoted as $\Pr(l_{ImageNet})$, where $l_{ImageNet}$ is the 1000 ImageNet categories. We use one round of 5-fold cross validation, and use all the images in urban tribe dataset for analysis.

We first check the relationshape of $\Pr(l_{urban})$ and $\Pr(l_{ImageNet})$ of candidate person images. We calculate the correlation coefficient of $\Pr(l_{urban})$ and $\Pr(l_{ImageNet})$ for all 1000 ImageNet classes, denoted as $R(\Pr(l_{urban}), \Pr(l_{ImageNet}))$. We also calculate the 1000 mutual information of the predicted score of urban tribe and ImageNet (where predicted score is 1 if the predicted label is the category being tested, 0 otherwise), denoted as $I(l_{urban}, l_{ImageNet})$.

In Figure 4, we choose two urban classes: *biker* and *hipster*, and plot the correlation coefficient R and mutual information I.The first row shows the result of *biker*, and second row *hipster*. For *biker*, there are several impulses in correlation plot, and one significant impulse in mutual information plot. *whiptail lizard* has high correlation with *biker*. Meanwhile, for *hipster*, which is the most difficult class, the correlation coefficient and mutual information are both low for all ImageNet classes.

To confirm the correlation,we use $\Pr(l_{ImageNet})$ directly as features, substituting the concatenated fc7fc6 features, and use the $\text{Nets}_{SDistort}$ approach for classification. The accuracy is 52.68%. This decent result indicates the relationship between l_{urban} and $l_{ImageNet}$. Then, we check $l_{ImageNet}$ with highest correlation coefficients with l_{urban}. In Figure 5, we choose four l_{urban}: *formal*, *ravers*, *goth*, *hipster*, and choose some examples of candidate person images that have both high $\Pr(l_{ImageNet})$ and high $\Pr(l_{urban})$. We also show examples of the images in $l_{ImageNet}$. We can see some of the shared features between corresponding person images and ImageNet images, for example, similar shape for women tops and stingrays (Figure 5b).

There is a correlation between class-wise accuracy of predicted l_{urban} and the degree of relationship between predicted $\Pr(l_{urban})$ and $\Pr(l_{ImageNet})$, as shown in Figure 6. Class-wise accuracy is calculated for candidate person images(Figure 6a, 6b) or scene images(Figure 6c, 6d). For each l_{urban}, the maximum correlation/mutual information over 1000 ImageNet classes are used to indicate the degree of its relationship with $l_{ImageNet}$.

The correlation between ImageNet class and urban tribe class and its relationship with class-wise recognition rate may indicate that the "generic" features extracted by pre-trained CNN networks are not so generic. The network is trained to separate the ImageNet classes most, and if we use the features for a new classification task, the performance of the task is related to how well the new classes can be "mapped" to the ImageNet classes.

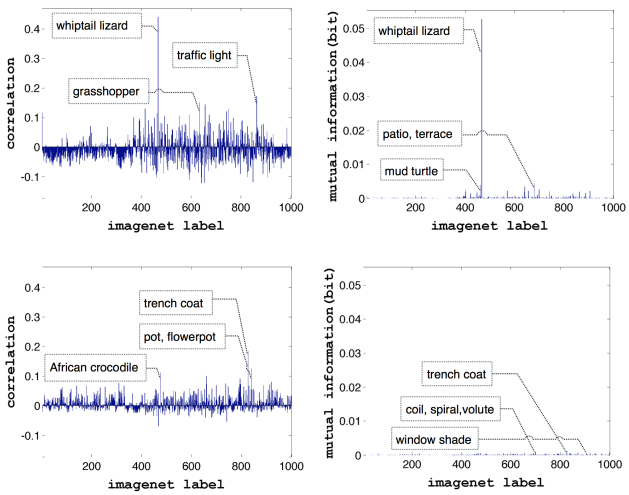


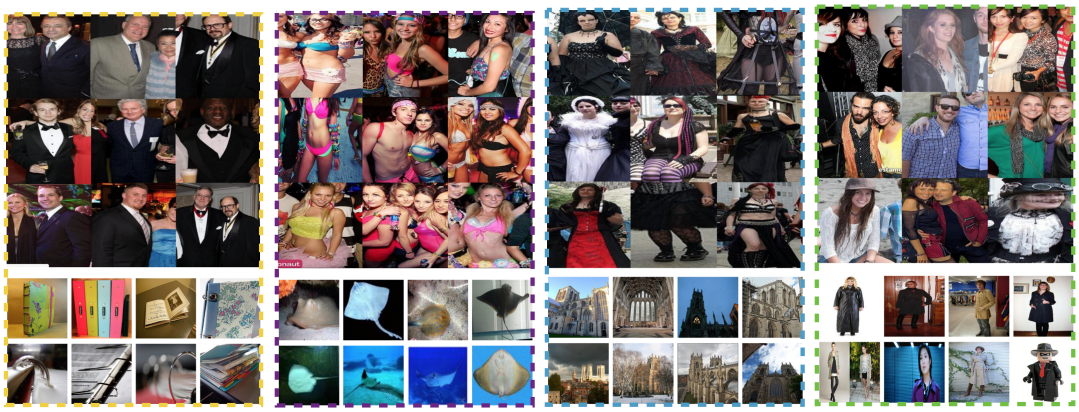
Figure 4: Correlation and mutual information of l_{urban} and 1000 $l_{ImageNet}$. The first row is for l_{urban}=Biker. The second row is for I_{urban}=Hipster. Top three ImageNet classes are marked.

5. Conclusion

In this work, we proposed a framework for social group recognition. The framework takes in both individual and global features. Features are extracted from fine-tuned CNN networks which has been pre-trained on ImageNet dataset, and then combined. Our results showed the success of our framework by achieving much better result than the previous work.

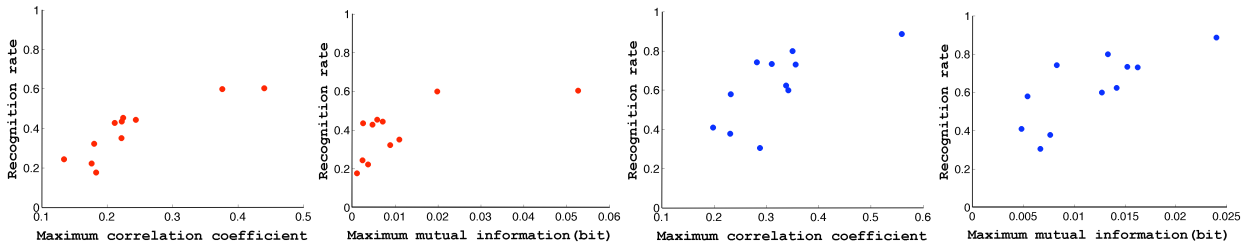
We also investigated into the pre-trained CNN features. Both visualization and numeric results showed the generalization ability of pre-trained CNN features to features of people's social styles, which has little shared features with the ImageNet object categories. Meanwhile, we found that there is a correlation between the probability of an image being in Imagenet classes and social group classes, and that better-recognized categories are more correlated with ImageNet categories.

In the future work, we intend to improve the classification performance by adapting convolutional networks more to social groups datasets. The relationship between ImageNet categories and urban tribes classes also brings forward an interesting future topic.



(a) Formal - Binder (b) Ravers - Stingray (c) Goth - Church (d) Hipster - Trench coat

Figure 5: selected urban tribe classes and the corresponding highest correlation $l_{ImageNet}$. Upper nine images: candidate person images with high $\Pr(l_{ImageNet})$ and high $\Pr(l_{urban})$. Lower eight images: example of images in $l_{ImageNet}$.



(a) Individual accuracy - maximum of R (b) Individual accuracy - maximum of I (c) Scene accuracy - maximum of R (d) Scene accuracy - maximum of I

Figure 6: The relationship between class-wise recognition rate and maximum of correlation $R(\Pr(l_{urban}), \Pr(l_{ImageNet}))$, class-wise recognition rate and maximum of mutual information $I(l_{urban}, l_{ImageNet})$

References

[1] I. S. Kwak, A. C. Murillo, P. Belhumeur, S. Belongie, and D. Kriegman, "From bikers to surfers: Visual recognition of urban tribes," in *British Machine Vision Conference (BMVC)*, (Bristol), September 2013.

[2] A. C. Murillo, I. S. Kwak, L. Bourdev, D. Kriegman, and S. Belongie, "Urban tribes: Analyzing group photos from a social perspective," in *CVPR Workshop on Socially Intelligent Surveillance and Monitoring (SISM)*, (Providence, RI), June 2012.

[3] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks.," *CoRR*, vol. abs/1311.2901, 2013.

[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.," *CoRR*, vol. abs/1310.1531, 2013.

[5] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks.," in *NIPS* (P. L. Bartlett, F. C. N. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, eds.), pp. 1106–1114, 2012.

[6] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural Computation*, vol. 1, pp. 541–551, 1989.

[7] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," in *Proceedings of the IEEE*, vol. 86(11), pp. 2278–2324, 1998.

[8] I. J. Goodfellow, D. Warde-Farley, M. Mirza, A. Courville, and Y. Bengio, "Maxout networks," 2013, arXiv:1302.4389.

[9] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. R. adn Dragomir Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going Deeper with Convolutions," 2014, arXiv:1409.4842.

[10] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "Overfeat: Integrated recognition, localization and detection using convolutional networks," in *International Conference on Learning Representations (ICLR2014)*, 2014.

[11] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic

segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[12] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," *CoRR*, vol. abs/1207.0580, 2012.

[13] I. J. Goodfellow, D. Warde-Farley, M. Mirza, A. Courville, and Y. Bengio, "Network in network," 2013, arXiv:1312.4400.

[14] S. Karayev, M. Trentacoste, H. Han, A. Agarwala, T. Darrell, A. Hertzmann, and H. Winnemoeller, "Recognizing Image Style," 2013, arXiv:1311.3715.

[15] P.-H. Gosselin, N. Murray, H. Jégou, and F. Perronnin, "Inria+xerox@fgcomp: Boosting the fisher vector for fine-grained classification," Tech. Rep. 0, INRIA, December 2013.

[16] Y. Jia, "Caffe: An open source convolutional architecture for fast feature embedding." http://caffe.berkeleyvision.org/, 2013.

[17] A. C. Gallagher and T. Chen, "Understanding images of groups of people.," in *CVPR*, pp. 256–263, IEEE, 2009.

[18] G. Wang, A. C. Gallagher, J. Luo, and D. A. Forsyth, "Seeing people in social context: Recognizing people and social relationships.," in *ECCV* (K. Daniilidis, P. Maragos, and N. Paragios, eds.), vol. 6215, pp. 169–182, Springer, 2010.

[19] A. Dhall, J. Joshi, I. Radwan, and R. Goecke, "Finding happiest moments in a social context.," in *ACCV 2* (K. M. Lee, Y. Matsushita, J. M. Rehg, and Z. Hu, eds.), vol. 7725, pp. 613–626, Springer, 2012.

[20] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "Liblinear: A library for large linear classification," *Journal of Machine Learning Research*, vol. 9, pp. 1871–1874, 2008.

[21] G. H. Laurens van der Maaten, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, November 2008.