Scene Recognition in User Preference Prediction Based on Classification of Deep Embeddings and Object Detection

Andrey Savchenko, Dr. of Science, Prof., Leading Researcher
Alexandr Rassadin, intern
Laboratory of Algorithms and Technologies for Network Analysis
National Research University Higher School of Economics (HSE), Nizhny Novgorod
Email: avsavchenko@hse.ru

July 11, 2019

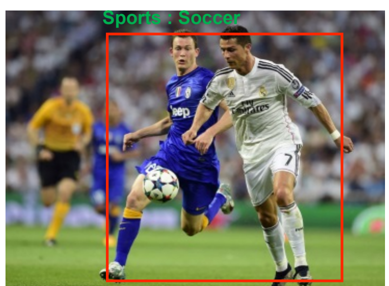
We introduce the outline of our talk

1. User preference prediction
2. Proposed approach
3. Experimental results
4. Concluding comments

Development of Preference Prediction Engine using Visual Data

– Deep understanding of user characteristics by analyzing user images and videos in a mobile device.

– Categorizing user's characteristics (taxonomy, demographics, hobbies, occupation, lifestyle, etc.) → Generate user profile



User Images : Categorized characteristics

User Profile

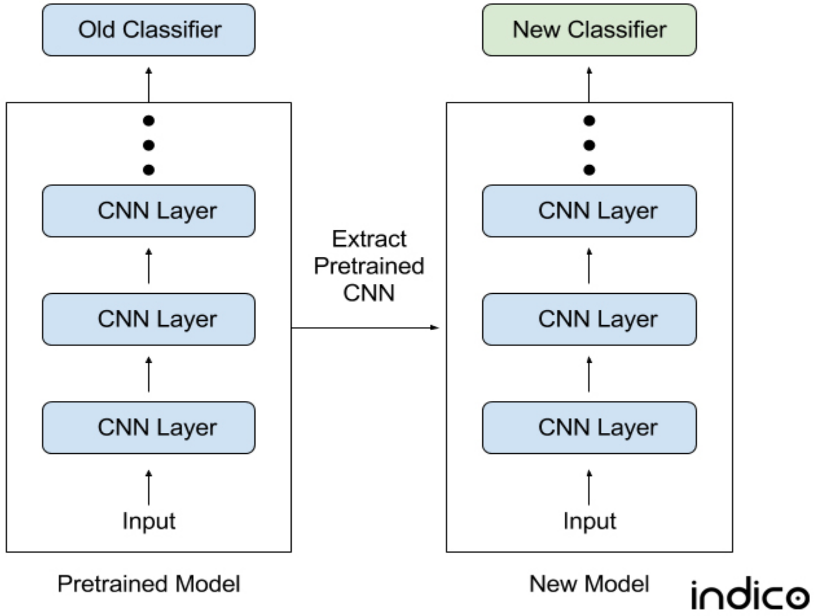
Image recognition

Problem formulation

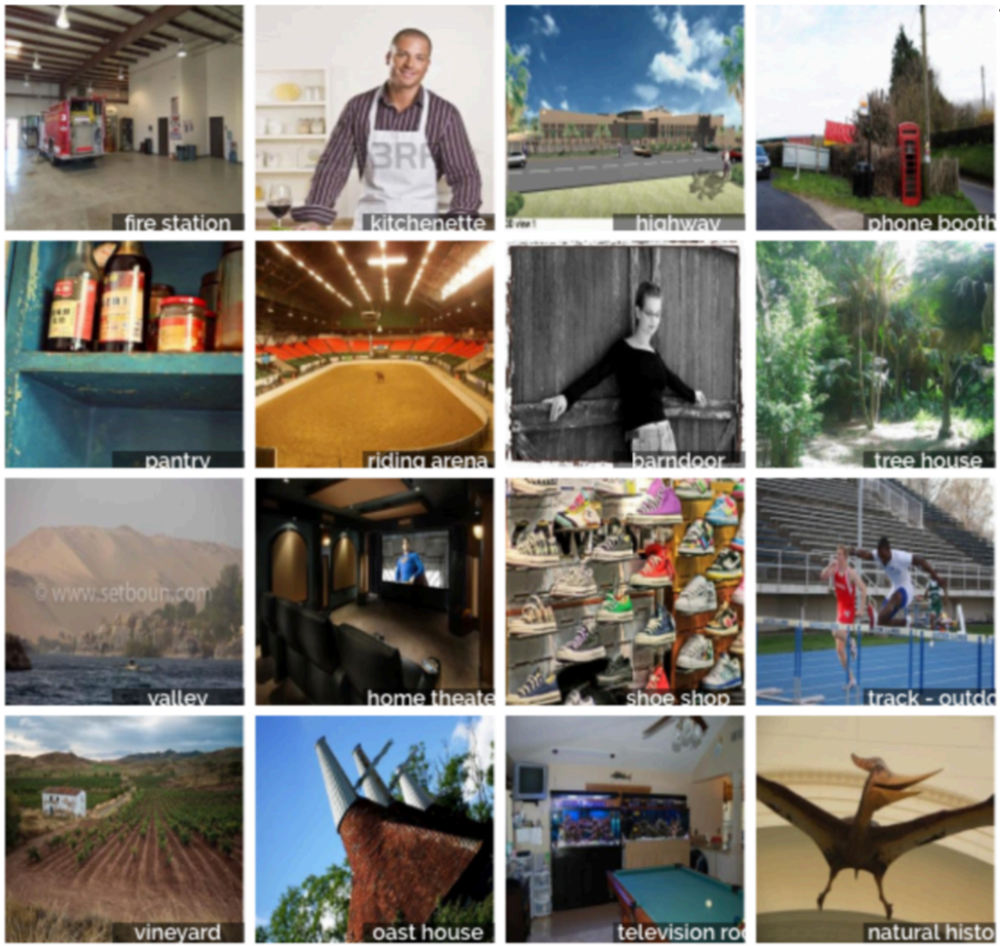
It is required to assign a new image X to one of C classes. Training set contains N reference images $\{X_n\}$, $n \in \{1,\ldots N\}$, with known class label $c_n \in \{1,\ldots C\}$

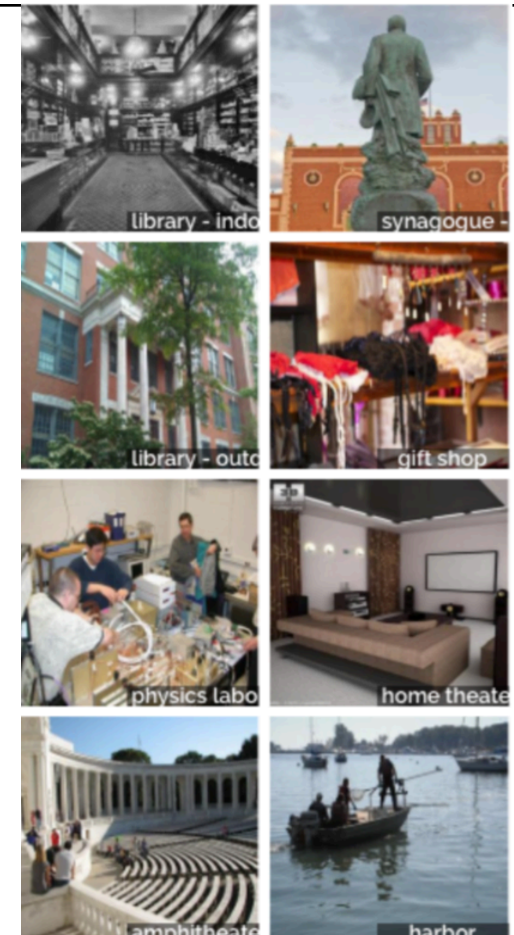
Conventional solution

Fine-tune convolutional neural network (CNN) pre-trained on ImageNet-1000



Scene recognition





Places2 scenes dataset, http://places2.csail.mit.edu

Event recognition

“An event captures the complex behavior of a group of people, interacting with multiple objects, and taking place in a specific environment. Images from the same event category may vary even more in visual appearance and structure” (Wang et al, IJCV 2018)

WIDER (Web Image Dataset for Event Recognition) dataset

Ensemble of three feature vectors

1. *Scores* of fine-tuned model: C-dimensional feature vector $\mathbf{p}=[p_1,\ldots, p_C]$
Training set is associated with scores $\{\mathbf{p}_n\}$, $\mathbf{p}_n=[p_{n,1},\ldots, p_{n;C}]$.

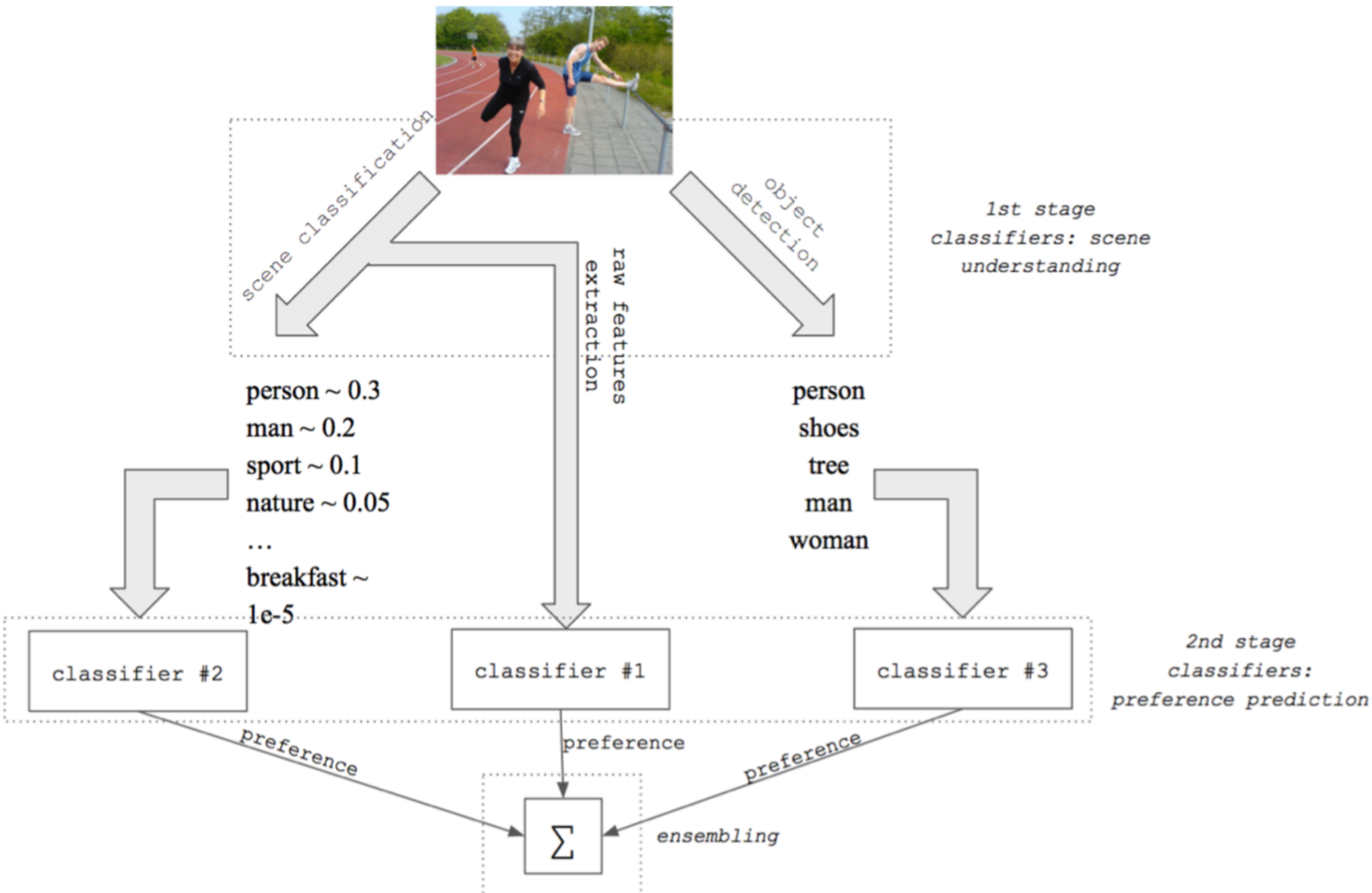
$$\sum_{c=1}^{C} p_c = 1$$

2. *Embeddings (features)* from either pre-trained or fine-tuned CNN: D-dimensional feature vector $\mathbf{x}=[x_1,\ldots, x_D]$
Training set is associated with embeddings $\{\mathbf{x}_n\}$, $\mathbf{x}_n=[x_{n,1},\ldots, x_{n;D}]$.

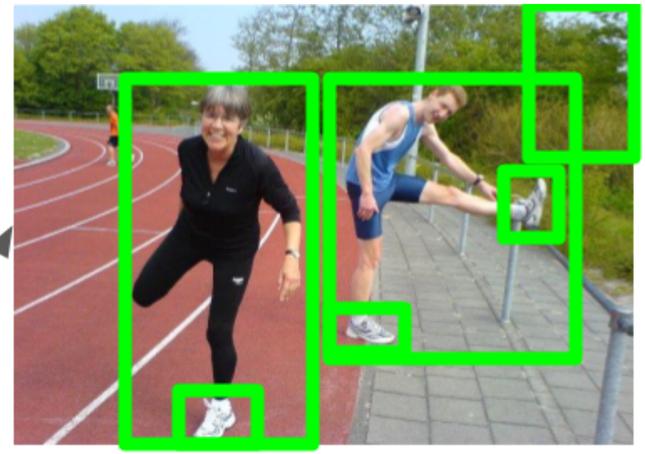
3. The scene is composed of parts and some of those parts can be named and correspond to objects

Object detection model predicts the positions of several objects in the input image and predict the scores of each class from the predefined set of $K > 1$ types. We ignore bounding boxes and extract the sparse vector $\mathbf{o} = [o_1,\ldots,o_K]$ of *detection scores* for each type. If there are several objects of the same type, the maximal score is stored in this feature vector.

Proposed pipeline



Example



person
shoes
tree
man
woman

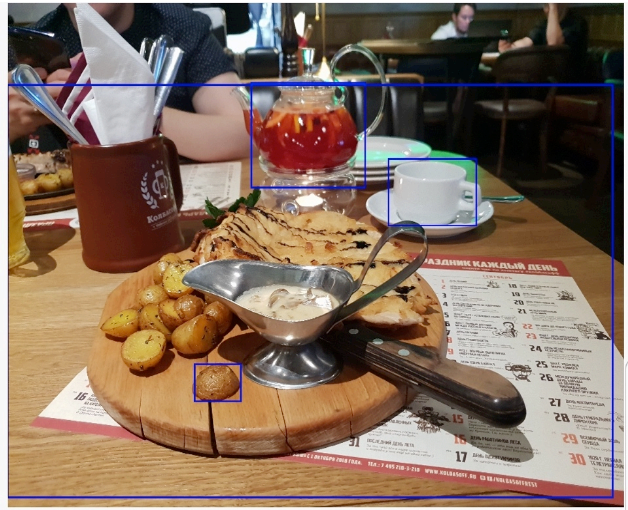
person ~ 0.3
man ~ 0.2
sport ~ 0.1
nature ~ 0.05
…
breakfast ~ 1e-5

Android demo application

PREV NEXT BACK

photo 67 out of 1196
Public photo
Scenes:60 ms
Москва, Россия

bus (0,92)
car (0,52)
scenes:street (0,38); parking garage (0,12);
No faces found
text:



PREV NEXT BACK

photo 360 out of 1196
Public photo
Scenes:62 ms
latitude=0,000 longitude=0,000

coffee cup (0,80)
food (0,48)
drink (0,36)
dining table (0,34)
scenes:restaurant (0,37);
No faces found

PREV NEXT BACK

photo 377 out of 1196
Public photo
Scenes:68 ms
latitude=0,000 longitude=0,000

equipment (0,44)
equipment (0,33)
scenes:football (0,70); athletic field (0,29);
No faces found
text:

Experiment 1. Subset of ImageNet dataset.

Dataset size: 45K

Number of classes C: 40 +1 distractor class from Caltech-101/256 datasets

sport scenes, drugstore, gas station, beauty shop/salon, spa, department store, bookstore, grocery store, amusement park, gallery, art gallery, picture gallery, music hall, opera, cinema, alehouse, cabaret, nightclub, night club, club, nightspot, news magazine, comic book

Experiment 1. Accuracy (%) of scene recognition models

	MobileNet v1	MobileNet v2 (α=1)	MobileNet v2 (α=1.4)	Inception v3
Fine-tuned CNN	76.84	78.66	80.02	82.22
Pre-trained features, FM (Factorization machine)	18.5	21.34	22.76	24.27
Pre-trained features, SVM	78.25	83.6	85.12	86.38
Fine-tuned scores, FM	24.11	27.8	28.92	29.0
Fine-tuned scores, SVM	72.76	76.25	77.9	77.91
Proposed ensemble, FM	48.35	50.8	52.56	53.0
Proposed ensemble, SVM	80.15	85.14	86.39	87.52

Factorization machines cannot improve the accuracy when compared to simple CNN-based scene recognition

Experiment 1. Out-of-class detection.

True negative

True positive

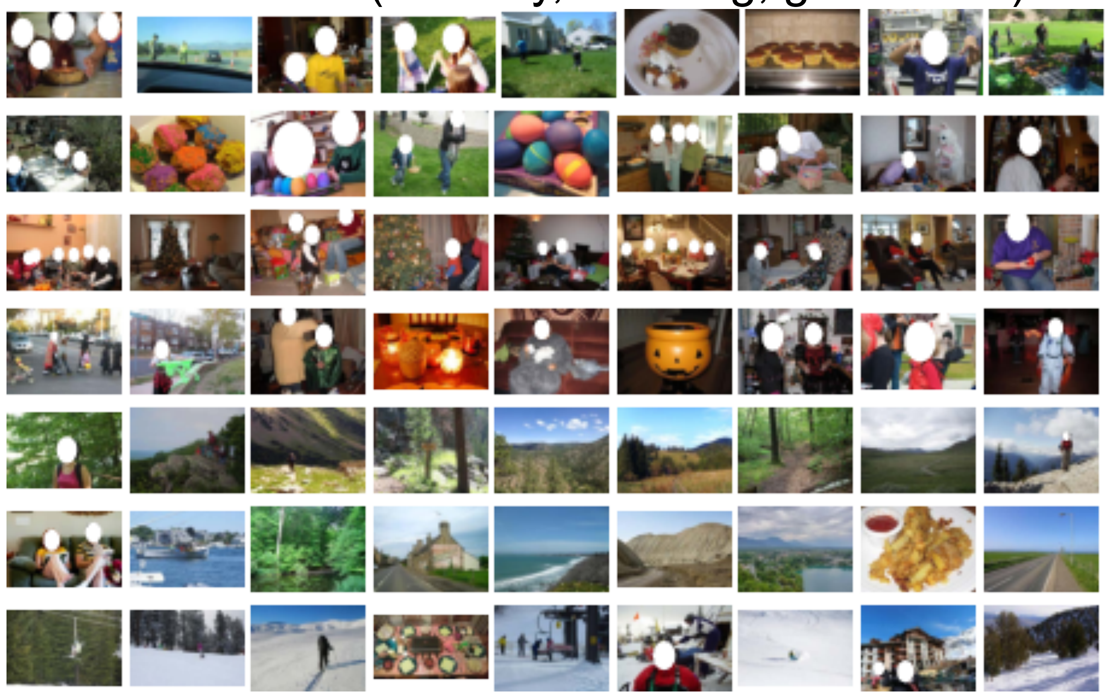
False positive (soccer scene is detected as out-of-class sample)

Experiment 2. Event recognition

1 PEC (Photo Event Collection) dataset

Dataset size: 61K

Number of classes (birthday, wedding, graduation) C: 14



2 WIDER dataset

Dataset size: 50K

Number of event classes (parade, dancing, meeting, press conference,…) C: 61

Experiment 2. Accuracy (%) for PEC dataset

Features	Classifier	Accuracy, %
MobileNet v2 ($\alpha = 1.4$), scores	Random Forest	56.20
	Linear SVM	51.95
	Fine-tuned	61.11
MobileNet v2 ($\alpha = 1.4$), features	Random Forest	57.09
	Linear SVM	58.32
	Fine-tuned	62.13
SSD+MobileNet	Random Forest	36.82
	Linear SVM	42.18
	Fine-tuned (new FC layer)	40.16
Our ensemble (client-side classifiers)	Random Forest	57.45
	Linear SVM	60.84
	Fine-tuned	63.34
Inception v3, scores	Random Forest	57.45
	Linear SVM	52.55
	Fine-tuned	61.81
Inception v3, features	Random Forest	58.31
	Linear SVM	61.82
	Fine-tuned	63.68
Faster R-CNN+InceptionResnet	Random Forest	44.59
	Linear SVM	48.83
	Fine-tuned (new FC layer)	47.45
Our ensemble (server-side classifiers)	Fine-tuned	64.98

We improved the previous state-of-the-art for PEC from 62.2% [Wang et al, IJCV 2018] even for client-side model (63.34%).

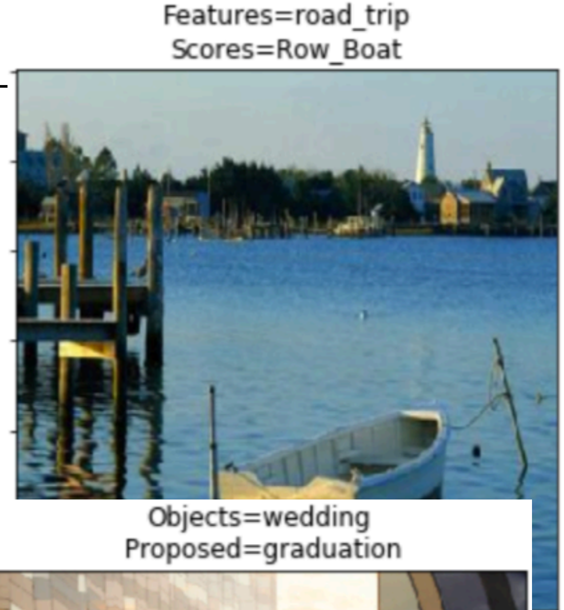
Our server-side model is better (accuracy 64.98%).

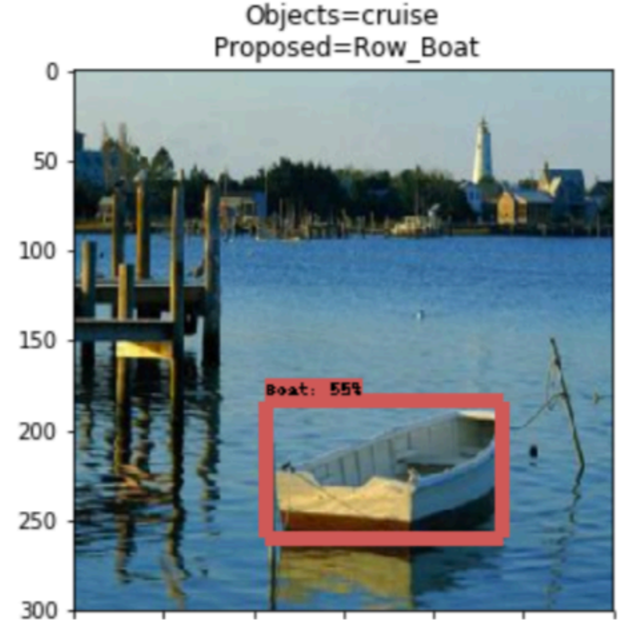
Experiment 2. Accuracy (%) for WIDER dataset

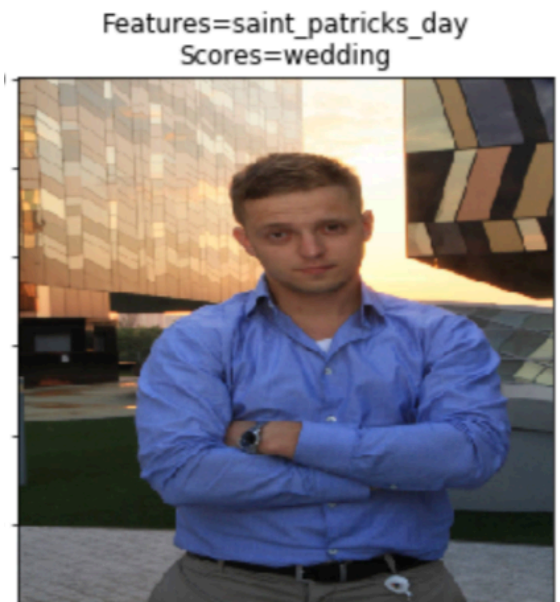
Features	Classifier	Accuracy, %
MobileNet v2 ($\alpha = 1.4$), scores	Random Forest	40.53
	Linear SVM	35.25
	Fine-tuned	40.49
MobileNet v2 ($\alpha = 1.4$), features	Random Forest	42.08
	Linear SVM	45.22
	Fine-tuned	49.48
SSD+MobileNet	Random Forest	15.91
	Linear SVM	19.91
	Fine-tuned	12.91
Our ensemble (client-side classifiers)	Fine-tuned	49.80
Inception v3, scores	Random Forest	41.61
	Linear SVM	34.91
	Fine-tuned	41.66
Inception v3, features	Random Forest	42.69
	Linear SVM	50.47
	Fine-tuned	50.96
Faster R-CNN+InceptionResnet	Random Forest	27.39
	Linear SVM	28.66
	Fine-tuned (new FC layer)	21.27
Our ensemble (server-side classifiers)	Fine-tuned	51.76

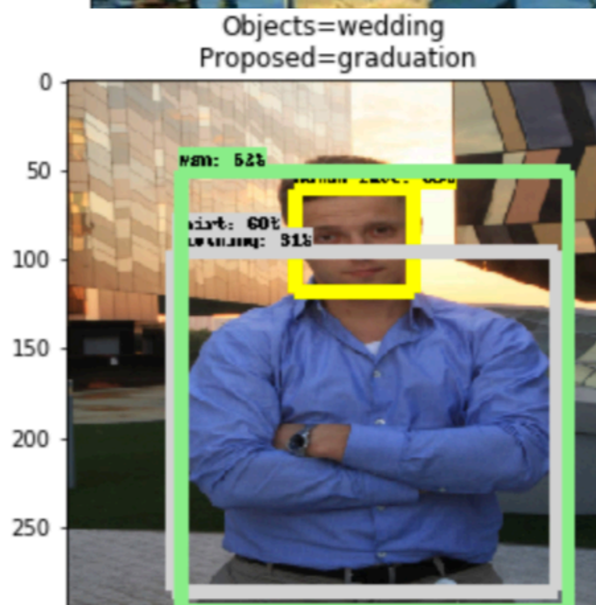
We have not still reached the state-of-the-art accuracy 53% [Wang et al, IJCV 2018]. However, our accuracy is 7.4-9.3% higher when compared to the best results (42.4%) from original paper (Xiong et al, CVPR 2015)

Experiment 2. Qualitative results









And summarizing our results we have the following conclusions

Proposed pipeline using fusion of classifiers has a list of advantages

1. It usually leads to the most accurate solution. We achieved state-of-the-art results for event recognition from Photo Event Collection dataset
2. Our approach was implemented in a special mobile application

And disadvantage

1. Slow processing especially if object detection is not needed. Simple MobileNet v2 with $\alpha = 1.4$ is the best choice for offline mobile applications with strict constraints to the running time
2. Still cannot obtain state-of-the-art accuracy for WIDER event collection

Future Works

1. Extend our solution for predicting the user preferences from a *set* of photos rather than process each photo independently
2. Improve the speed by using fast object detectors, approximate NN search, structural pruning of CNNs, etc.

Thank you for your attention

Any Questions?