NATIONAL RESEARCH UNIVERSITY

National Research University Higher School of Economics (HSE University) – N. Novgorod

EFFICIENT IMAGE RECOGNITION WITH MULTI-TASK NEURAL NETWORKS

Andrey V. Savchenko
Dr. of Sci., Prof.,
Lead Researcher in HSE's international laboratory LATNA
Email: avsavchenko@hse.ru
URL: www.hse.ru/en/staff/avsavchenko

Huawei On-device Artificial Intelligence Workshop
October 30, 2020

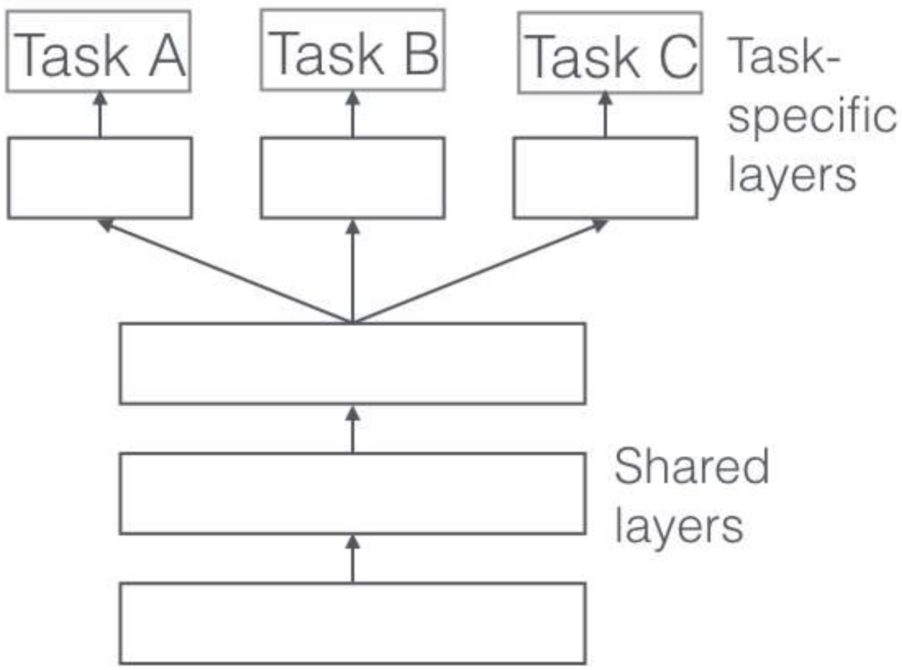
- Memory restrictions
- Computational complexity
- Power consumption

Multiple tasks should be solved

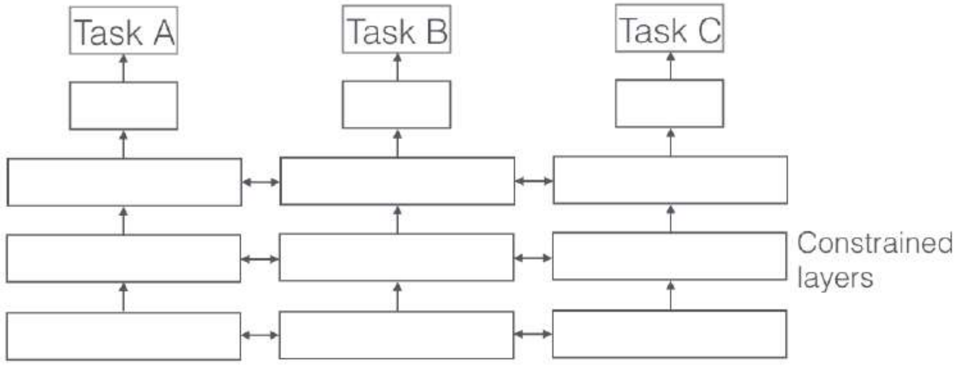
↓

Multi-task learning

Hard parameter sharing



Soft parameter sharing



https://ruder.io/multi-task/

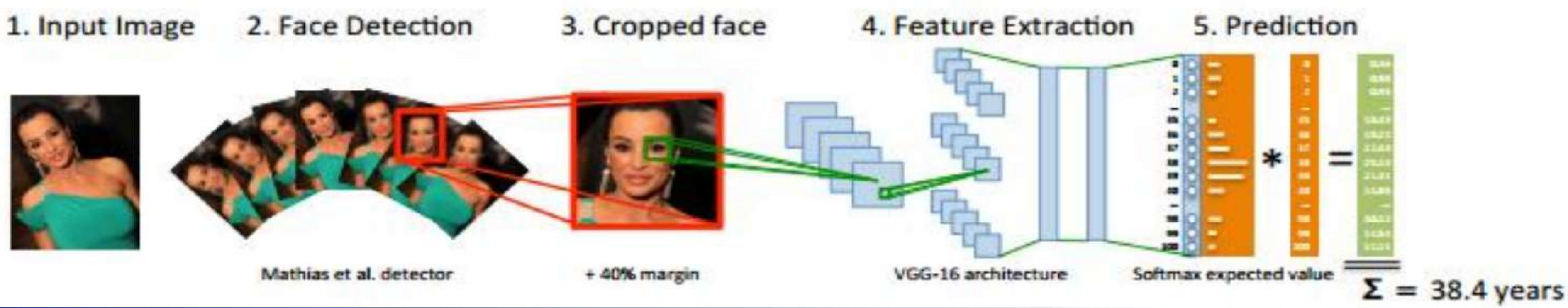
1. Facial attribute recognition
2. Understanding advertisements
3. Scene and event recognition
4. Food classification and restaurant recommendation

Facial attribute recognition

Adience

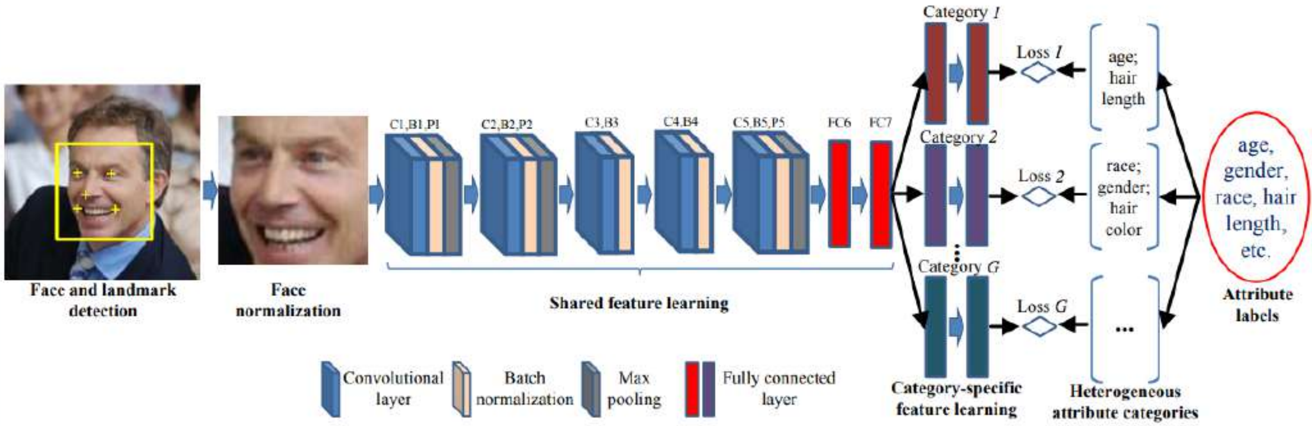
IMDB/Wiki

DEX: Deep EXpectation of apparent age from a single image, 2015 (VGG16-Net)

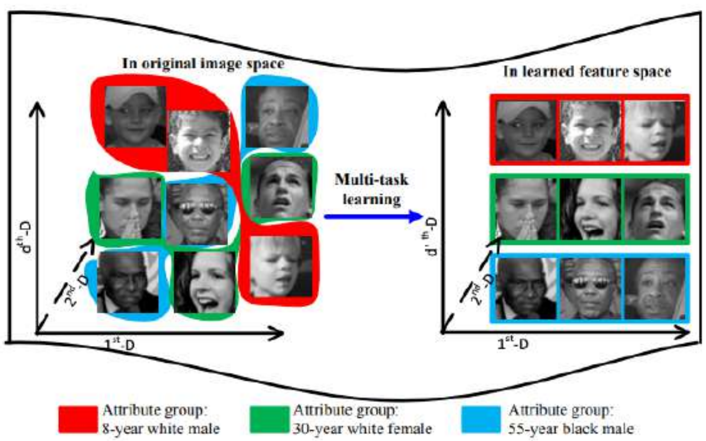


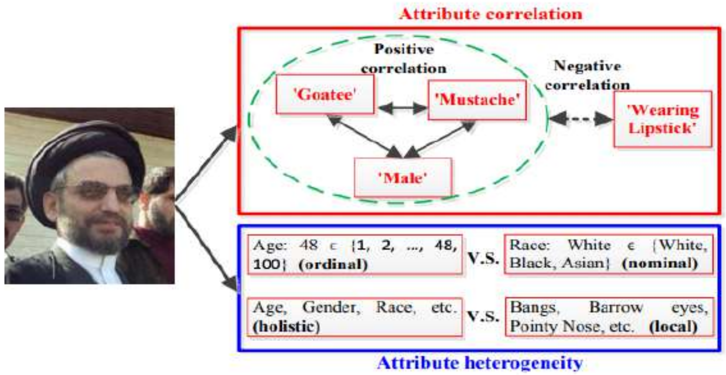
Multi-task attribute recognition (1)

$$\arg\min_{W_c,\{W^j\}_{j=1}^{M}} \sum_{j=1}^{M}\sum_{i=1}^{N}\mathcal{L}(y_i^j, \mathcal{F}(X_i, W^j \circ W_c)),$$
$$+\gamma_1\Phi(W_c) + \gamma_2\Phi(W^j)$$



The same descriptors make different classes more distinguishable





Han et al Heterogeneous Face Attribute Estimation: A Deep Multi-Task Learning Approach, PAMI 2017

Multi-task attribute recognition (2)

AffectNet

Algorithm 1: Packing and Expanding (PAE)

Input: given task 1 and an original model trained on task 1.

Set an accuracy goal for task 1;

Alternatingly remove small weights and re-train the remaining weights for task 1 via iterative pruning [29], until meeting the accuracy goal;

Let the model weights reserved for task 1 be W_1 (referred to as task-1 weights), and those that are removed by the iterative pruning be W_1^r (referred to as the saved weights);

for *task* $i = 2 \cdots K$ *(let the saved weights of task* i *be* W_{i-1}^r*)* **do**

  Set an accuracy goal for task i;

  Use the weights W_1 and W_{i-1}^r to train task i, with W_1 fixed;

  If the accuracy goal is not achieved by the trained model, expand the number of filters (wights) in the model, and reset $W_{i-1}^r \leftarrow W_{i-1}^r \cup W_E$, where W_E denotes the expanded weights;

  Alternatingly remove small weights from W_{i-1}^r and re-train the remaining weights (with W_1 fixed) for task i via iterative pruning, until meeting the accuracy goal;

end

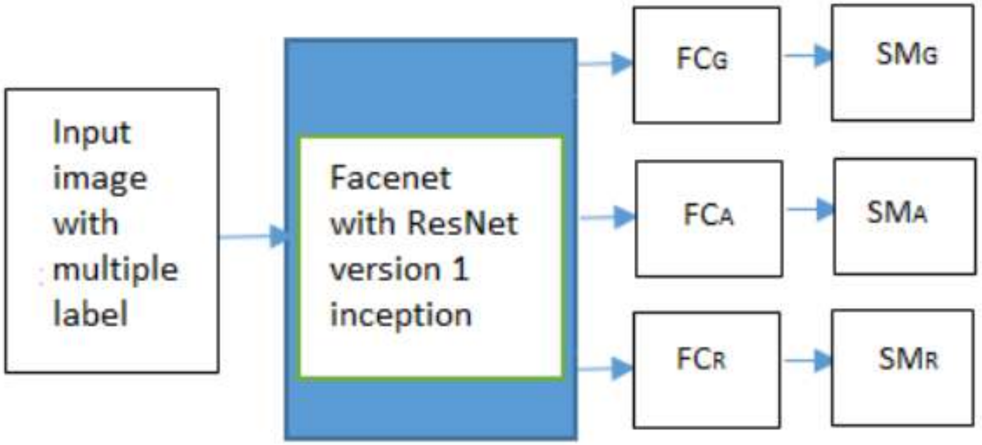
Multiple tasks:

- face verification (99.67% on LFW)
- age prediction (57.30% on Adience)
- gender classification (92.23% on Adience)
- emotion recognition (65.29% on AffectNet)

http://mohammadmahoor.com/affectnet/
Steven C. Y. Hung et al, ICMR 2019

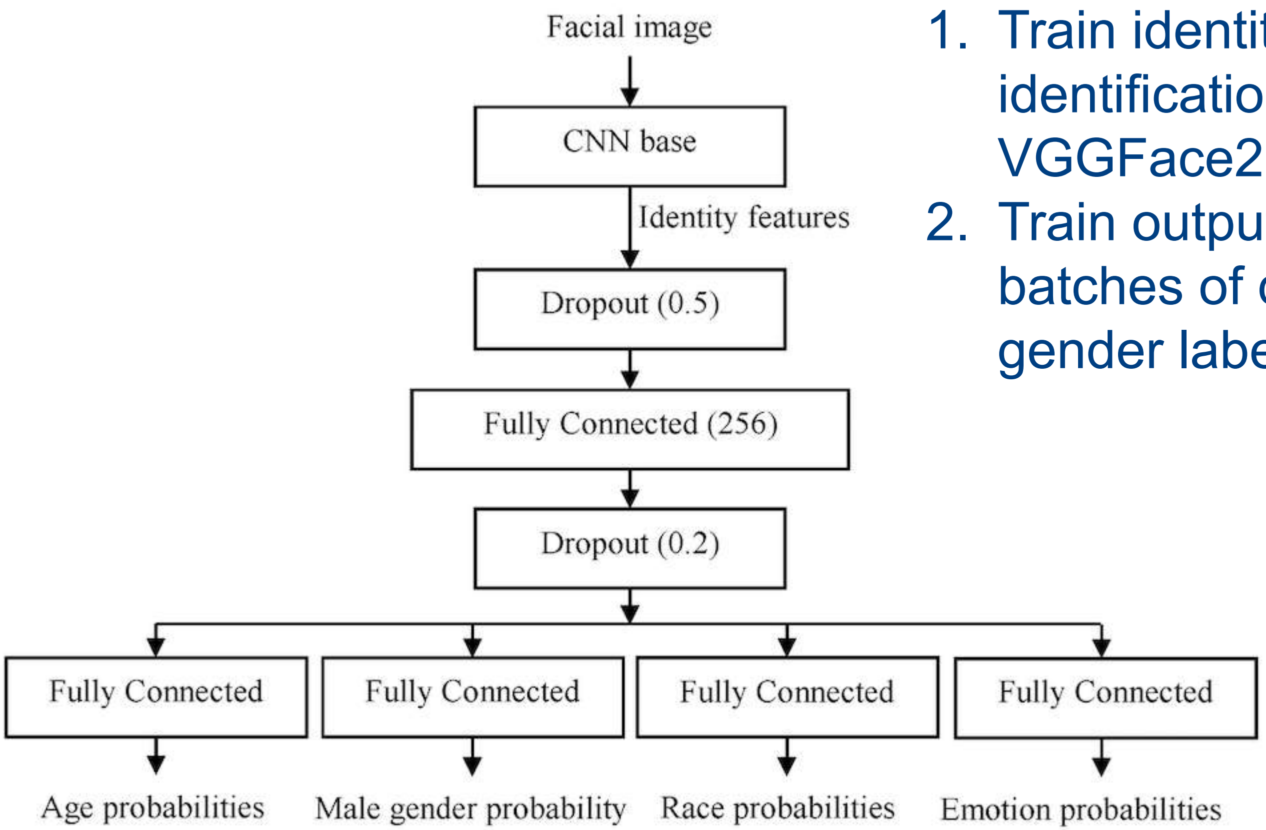
Age/gender/race recognition

UTKFace dataset



Das et al Mitigating Bias in Gender, Age and Ethnicity Classification, ECCV 2018

Multi-output model for age, gender and identity recognition



1. Train identity features for face identification (softmax loss, VGGFace2 dataset)
2. Train outputs by alternating batches of data with age and gender labels

Age prediction

$$l^* = \frac{\sum_{k=1}^{K} l_{(k)} \cdot P\left(l_{(k)} \middle| X(t)\right)}{\sum_{k=1}^{K} P\left(l_{(k)} \middle| X(t)\right)}.$$

$$P\left(l_{(1)} \middle| X(t)\right) \geq P\left(l_{(2)} \middle| X(t)\right) \geq \ldots \geq P\left(l_{(L)} \middle| X(t)\right)$$

[Savchenko, PeerJ Computer Science, 2019]
https://github.com/HSE-asavchenko/HSE_FaceRec_tf/tree/master/age_gender_identity

Experimental results: Age/gender recognition for UTKFace (In-the-wild faces) dataset

Models	Gender accuracy, %	Age MAE	Age accuracy, %	Model size, Mb	Inference time, ms
DEX	91.05	6.48	51.77	1050.5	47.1
Wide ResNet (weights.28-3.73)	88.12	9.07	46.27	191.2	10.5
Wide ResNet (weights.18-4.06)	85.35	10.05	43.23	191.2	10.5
FaceNet	89.54	8.58	49.02	89.1	20.3
BKNetStyle2	57.76	15.94	23.49	29.1	12.5
SSRNet	85.69	11.90	34.86	0.6	6.6
MobileNet v2 (Agegendernet)	91.47	7.29	53.30	28.4	11.4
ResNet-50 from InsightFace	87.52	8.57	48.92	240.7	25.3
"New" model from InsightFace	84.69	8.44	48.41	1.1	5.1
Inception trained on Adience	71.77	-	32.09	85.4	37.7
age_net/gender_net	87.32	-	45.07	87.5	8.6
MobileNets with single head	93.59	5.94	60.29	25.7	7.2
Proposed MobileNet, fine-tuned from ImageNet	91.81	5.88	58.47	13.8	4.7
Proposed MobileNet, pre-trained on VGGFace2	93.79	5.74	62.67	13.8	4.7
Proposed MobileNet, fine-tuned	94.10	5.44	63.97	13.8	4.7

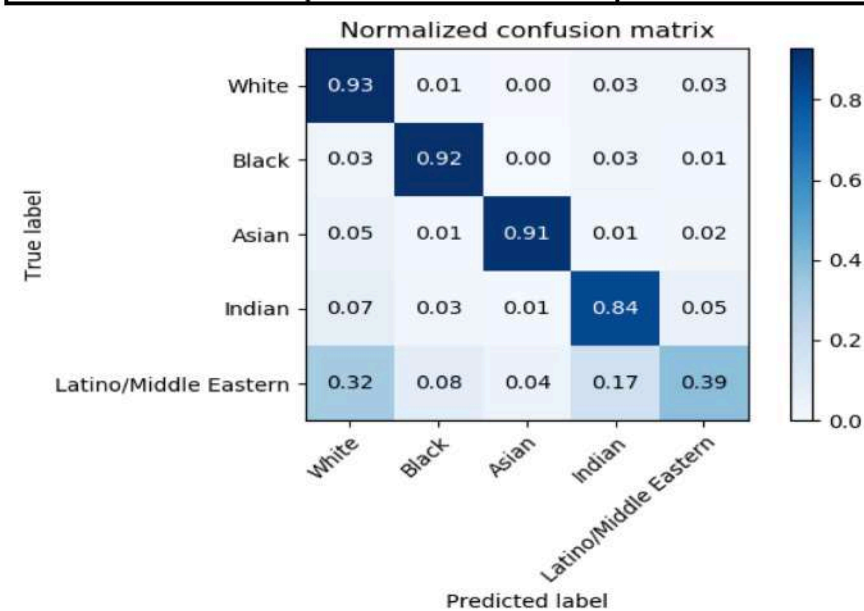
Face extraction and alignment: https://github.com/dandynaufaldi/Agendernet

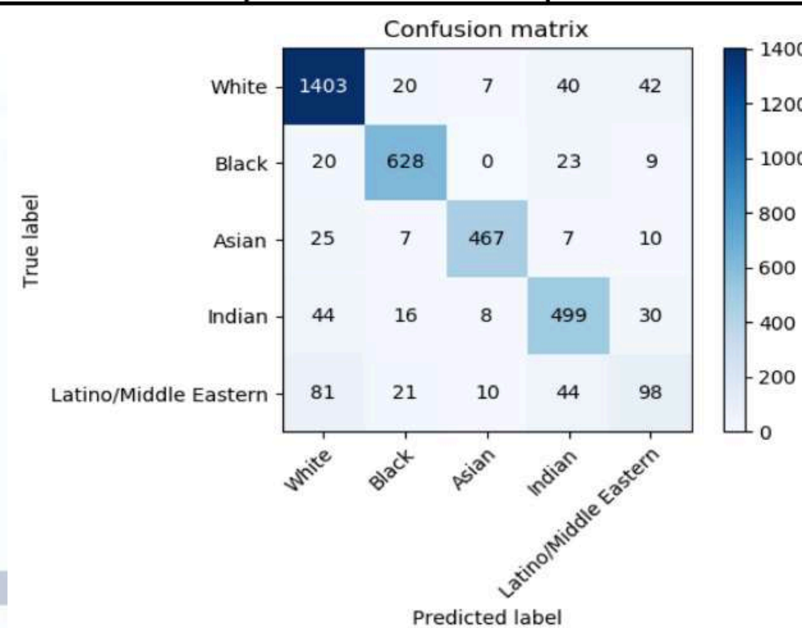
Experimental results: Age/gender recognition for UTKFace (aligned & cropped faces)

Models	Gender accuracy, %	Age MAE	Age accuracy, %
DEX	83.16	9.84	41.22
Wide ResNet (weights.28-3.73)	73.01	14.07	29.32
Wide ResNet (weights.18-4.06)	69.34	13.57	37.23
FaceNet	86.14	9.60	44.70
BKNetStyle2	60.93	15.36	21.63
SSRNet	72.29	14.18	30.56
MobileNet v2 (Agegendernet)	86.12	11.21	42.02
ResNet-50 from InsightFace	81.15	9.53	45.30
"New" model from InsightFace	80.55	8.51	48.53
Inception trained on Adience	65.89	-	27.01
age_net/gender_net	82.36	-	34.18
MobileNets with single head	91.89	6.73	57.21
Proposed MobileNet, fine-tuned from ImageNet	84.30	7.24	58.05
Proposed MobileNet, pre-trained on VGGFace2	91.95	6.00	61.70
Proposed MobileNet, fine-tuned	91.95	5.96	62.74

Ethnicity recognition results, UTKFace

	VGGFace	VGGFace-2	FaceNet	Our MobileNet	
				Age/gender features	Identity features
Random Forest	83.5	87.8	84.3	80.1	83.8
k-NN	76.2	84.5	84.4	72.2	82.2
SVM (RBF)	78.8	82.4	86.2	82.8	87.7
Linear SVM	79.5	83.1	85.6	80.6	85.6
New Dense layer	80.4	86.4	84.4	80.1	87.0





Android demo app

PREV NEXT BACK

photo 21 out of 136
Private photo
Selfie
latitude=0,000 longitude=0,000

no objects found
scenes:lecture/conference (0,28); nursing home (0,22);
child 1: age=8 male
girl friend: age=34 female
me: age=28 male
age=8 female
child 2: age=3 male
age=9 male
age=6 female
age=34 female
age=3 male
text:

100%

PREV NEXT BACK

photo 47 out of 246
Private photo
latitude=0,000 longitude=0,000

footwear (0,39)
scenes:beach (0,41); boardwalk (0,29);
age=25 female white
age=38 male black
age=27 female white
child 2: age=2 male white
age=34 male black
text:

100%

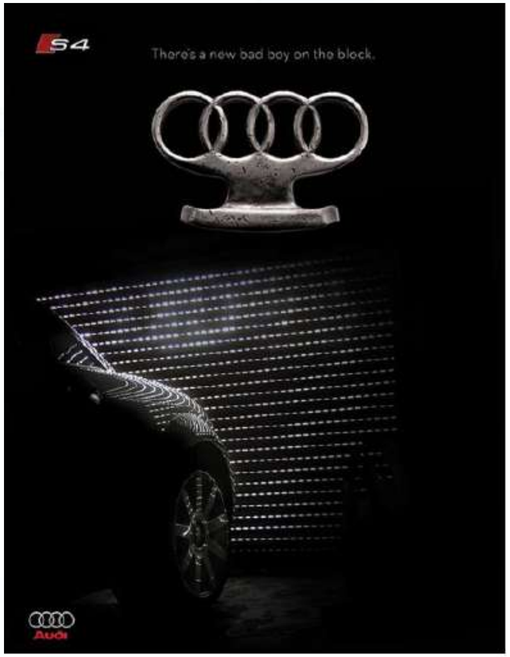
PREV NEXT BACK

photo 285 out of 1295
Public photo
latitude=0,000 longitude=0,000

no objects found
scenes:stage (0,71);
age=33 female asian
age=25 female asian
age=24 female asian
age=13 female asian
text:

100%

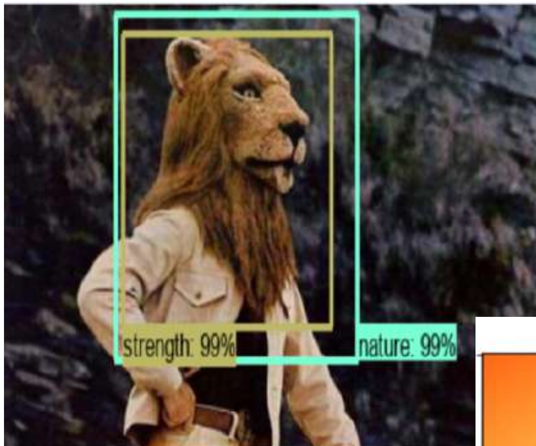
Understanding advertisements



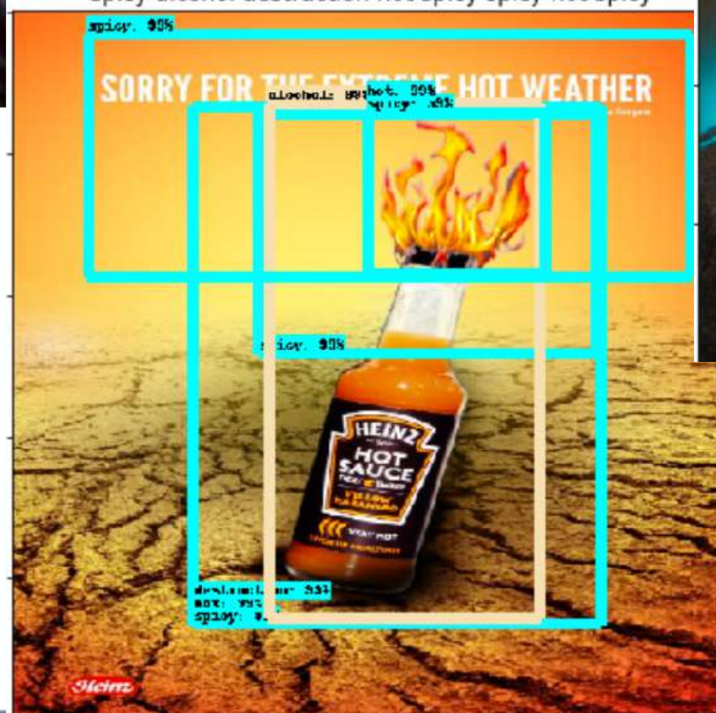
WWF anti-deforestation ad
Topic: environment.
Sentiment: alarmed.
Symbols: environment.

Audi ad: a new bad boy on the block
Topic: cars
Sentiment: inspired
Symbols: n/a.

http://people.cs.pitt.edu/~kovashka/ads/

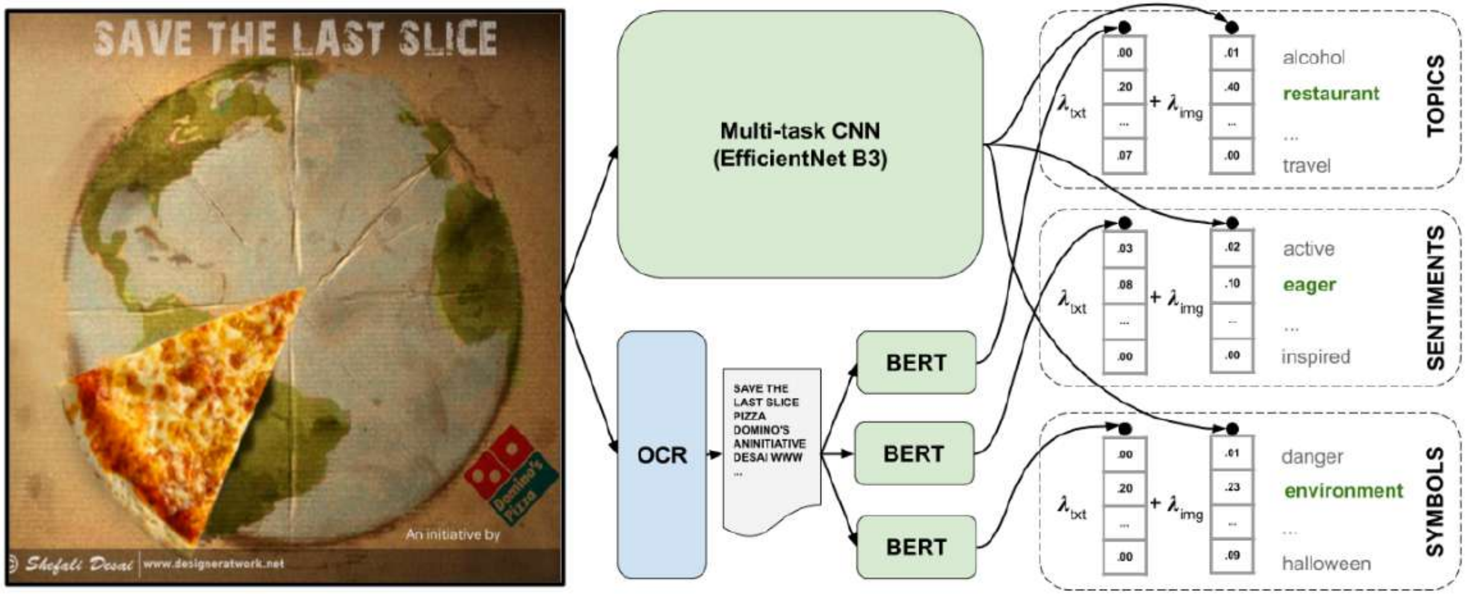


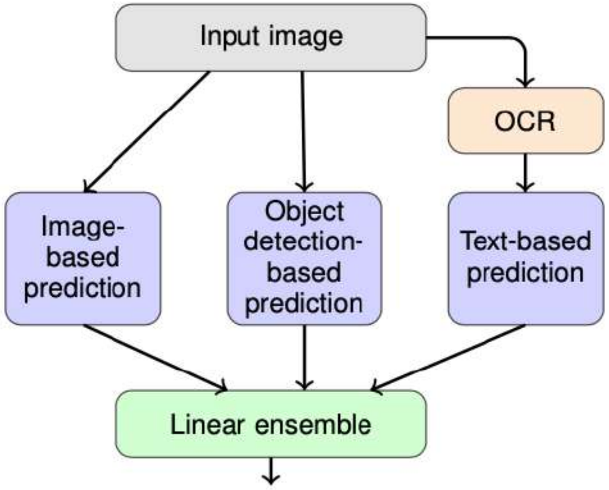
spicy alcohol destruction hot spicy spicy hot spicy



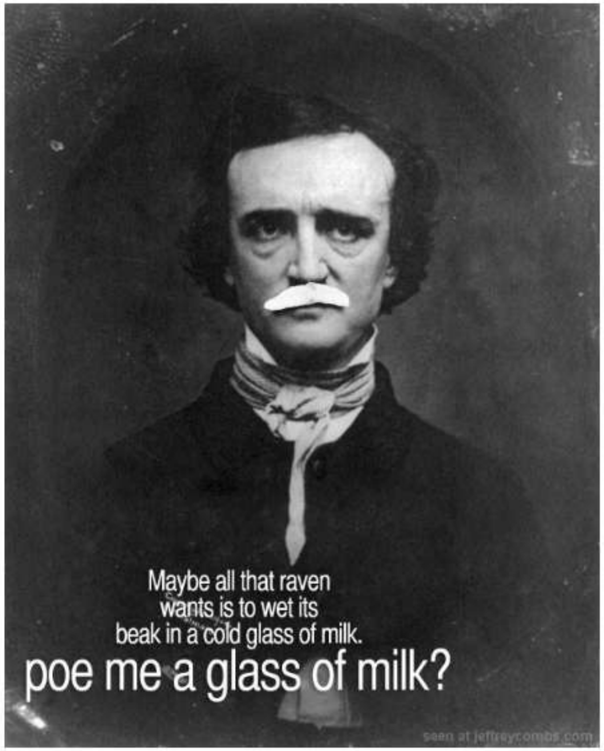
danger accident dangerous accident injury harm accident

Pipeline





Savchenko et al, COLING 2020

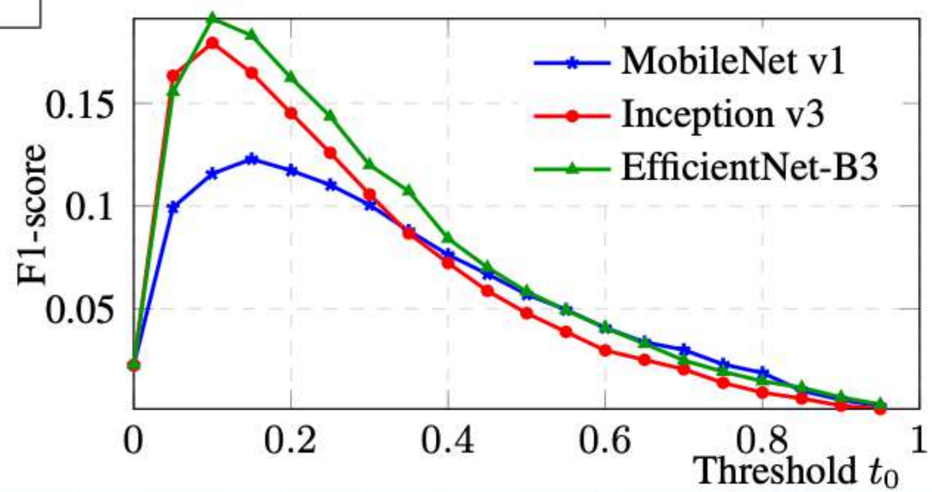


Tesseract Smith (2007)	Maybe all that raven wants is to wet its beak in a'cold glass of milk. poe mea lo EISsTo) aU eg
EAST+*Tesseract* Kopeykina and Savchenko (2019)	Maybe are Gein) eu SH Hostel S to wet its ake cold Me TS 10a milk. eee me glass 0) aa Penal see
PSENet Wang et al. (2019)	Elle that eV WM eMey iy is 10 Wed Mey in beak — Ey milk. of (eek glass — ranlll@a a me al eee) glass be at jeffreycombs-com
EasyOCR JaidedAI (2020)	Maybe all that raven poe me a glass ofmik? beak in a cold glass ofmik. wants is to wet its seen 3t jefieycomlzesolm
Charnet Xing et al. (2019)	MAYBE ALL THAT RAVEN WANTS WET ITS BEAKIN COLD GLASS MILK POE GLASS MILK? SEEN ATJ COM
CloudVision Otani et al. (2018)	Maybe all that raven wants is to wet its beak in a cold glass of milk. poe me a glass of milk? seen at jeffreycombs.com

Topics/sentiments

CNN	Topics	Sentiments
Baseline (Hussain et al., 2017)	60.34	27.92
ResNet-50	53.90	34.34
Resnet-152	52.67	27.58
Resnet-152 V2	52.12	27.64
MobileNet v1	50.56	33.50
MobileNet v2	54.76	**34.58**
EfficientNet-B0	60.06	34.03
EfficientNet-B3	**62.62**	34.12
Our multitask model	**62.99**	**36.27**

Symbols



Blending with OCR, symbols

OCR	Model	Text-based (w/texts) Acc./$F1_{micro}$	Text-based (w/texts) $F1_{macro}$	Blend (backoff) Acc./$F1_{micro}$	Blend (backoff) $F1_{macro}$
Multilabel symbol classification, 221 labels.		Image-based results: $F1_{micro}$ 0.1928, $F1_{macro}$ 0.1025.			
Tesseract	Bag-of-NGrams	0.0881	0.0393	0.1933	0.0956
	BERT	0.0220	0.0218	0.1942	0.1010
	RoBERTa	0.0220	0.0218	0.1936	0.1021
EAST+T	Bag-of-NGrams	0.1198	0.0559	0.2076	0.1023
	BERT	0.0952	0.0150	0.1945	0.0985
	RoBERTa	0.0225	0.0223	0.1939	0.0985
PSENet	Bag-of-NGrams	0.1172	0.0620	0.2020	0.1031
	BERT	0.0990	0.0154	0.1961	0.0995
	RoBERTa	0.0226	0.0224	0.1946	0.1002
Charnet	Bag-of-NGrams	0.1684	0.0967	0.2156	0.1146
	BERT	0.1354	0.0203	0.1964	0.0998
	RoBERTa	0.1441	0.0253	0.1974	0.0995
	MMBT (orig.)	–	–	0.0962	0.0671
	MMBT (upd.)	–	–	0.1078	0.0757
Google Cloud Vision	Bag-of-NGrams	**0.1830**	**0.1060**	**0.2249**	**0.1175**
	BERT	0.1520	0.0252	0.1968	0.1014
	RoBERTa	0.1580	0.0263	0.2017	0.1004
	MMBT (orig.)	–	–	0.1202	0.0825
	MMBT (upd.)	–	–	0.1099	0.0812

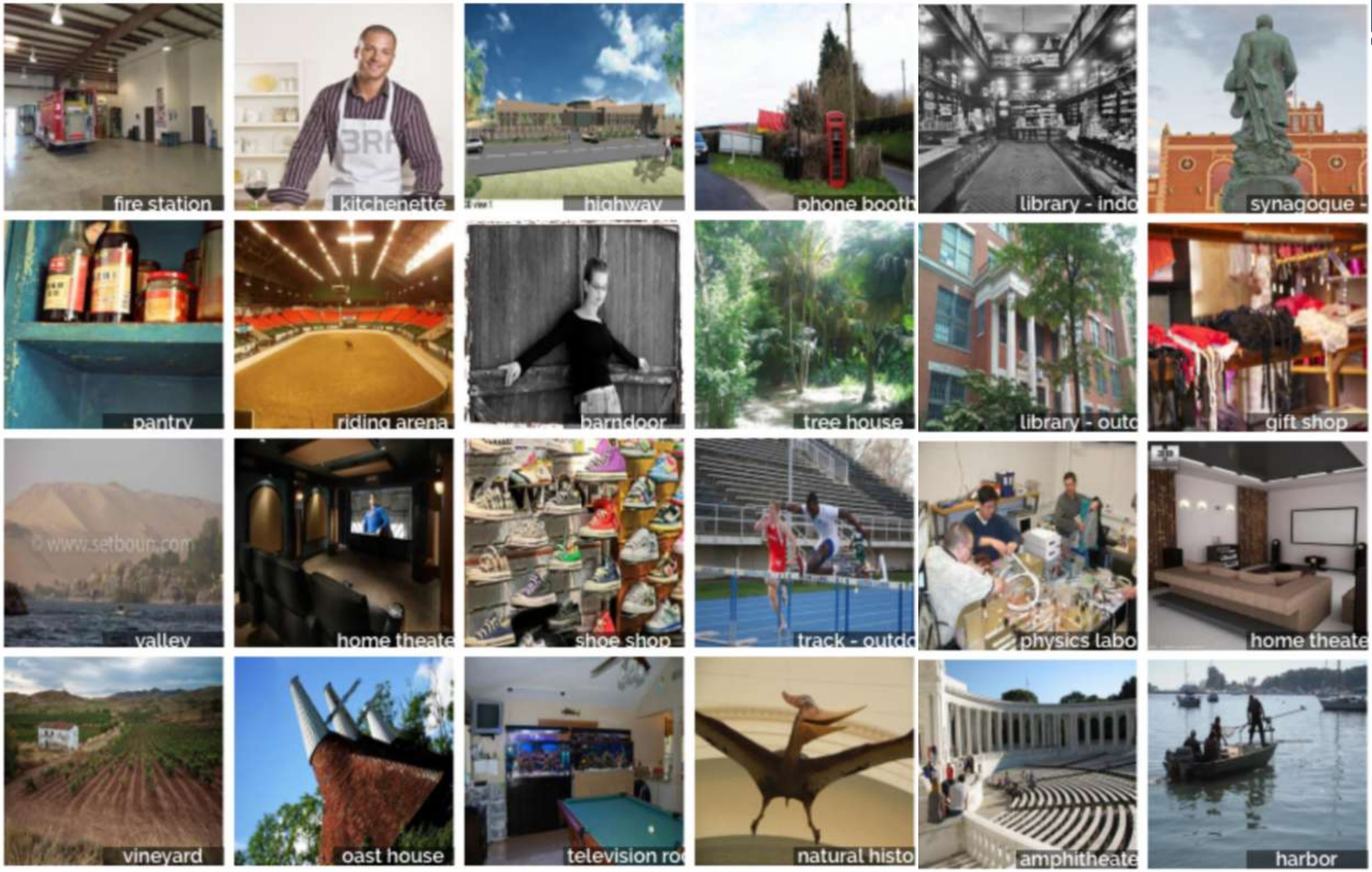
Blending with OCR, topics/sentiments

Topic classification.	Image-based results: accuracy 0.6299, $F1_{macro}$ 0.3800.				
Google Cloud Vision	Bag-of-Ngrams	0.6391	0.4531	0.7227	0.4840
	BERT	**0.7149**	**0.5573**	0.7599	0.5736
	RoBERTa	0.7109	0.5492	0.7557	0.5623
	MMBT (orig.)	–	–	0.7031	0.5396
	MMBT (upd.)	–	–	**0.7686**	0.5700
Charnet	Bag-of-Ngrams	0.6340	0.4502	0.7213	0.4816
	BERT	0.6985	0.5515	0.7536	**0.5793**
	RoBERTa	0.6933	0.5473	0.7545	0.5722
	MMBT (orig.)	–	–	0.6821	0.5357
	MMBT (upd.)	–	–	0.7534	0.5441

Sentiment classification.	Image-based results: accuracy 0.3627, $F1_{macro}$ 0.1041.				
Google Cloud Vision	Bag-of-Ngrams	0.2641	0.0859	0.3676	0.1061
	BERT	0.2595	0.1023	**0.3731**	0.1117
	RoBERTa	0.2750	0.1165	0.3697	0.1072
	MMBT (orig.)	–	–	0.3152	0.0925
	MMBT (upd.)	–	–	0.3224	**0.1219**
Charnet	Bag-of-Ngrams	0.2705	0.0905	0.3675	0.1062
	BERT	0.2497	0.1000	0.3675	0.1093
	RoBERTa	**0.2774**	**0.1211**	0.3717	0.1093
	MMBT (orig.)	–	–	0.2836	0.1049
	MMBT (upd.)	–	–	0.3053	0.1141

Scene and event recognition

Scene recognition



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

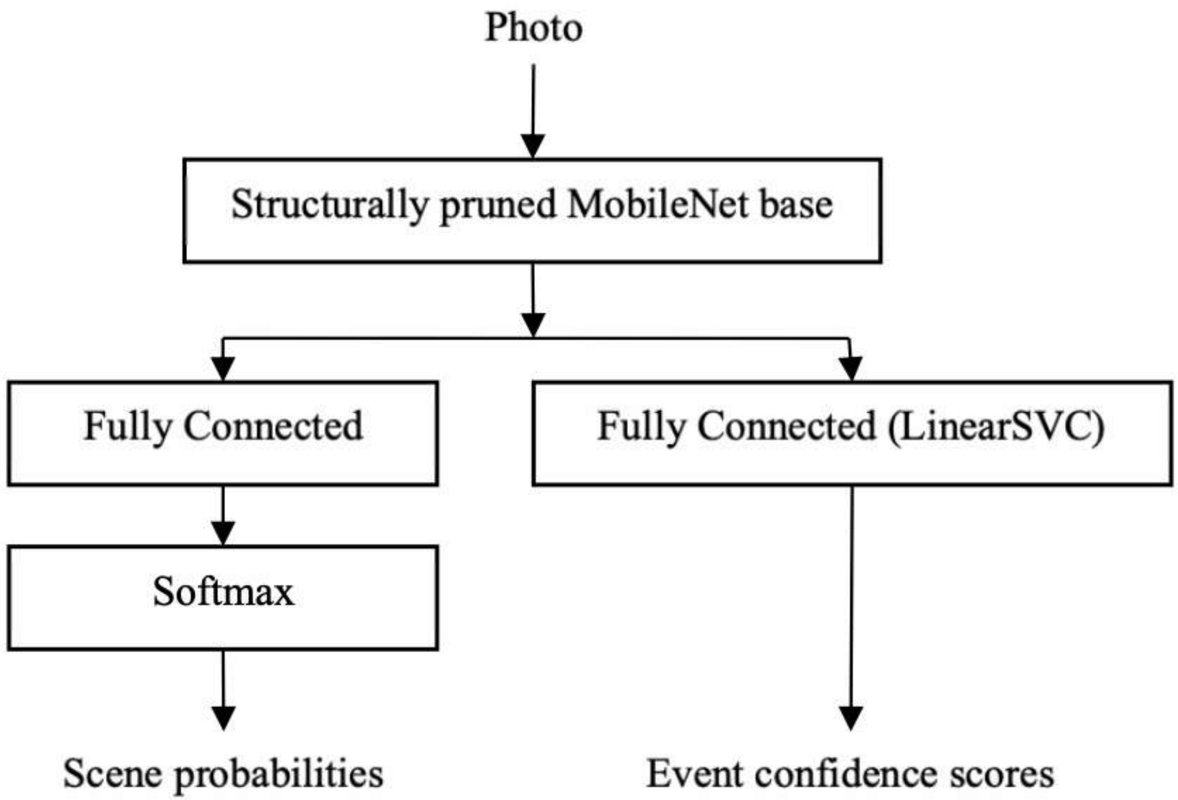
"An event captures the complex behavior of a group of people, interacting with multiple objects, and taking place in a specific environment. Images from the same event category may vary even more in visual appearance and structure" (Wang et al, IJCV 2018)

WIDER (Web Image Dataset for Event Recognition)

PEC (Photo Event Collection)

Children's birthday
Easter
Christmas
Halloween
Hiking
Road Trip
Skiing

Image recognition: it is required to assign an observed image X to one of C classes. Training set contains N reference images (examples) $\{X_n\}$, $n \in \{1,\dots N\}$, with known class label $c_n \in \{1,\dots C\}$



Savchenko A.V. IJCNN 2020

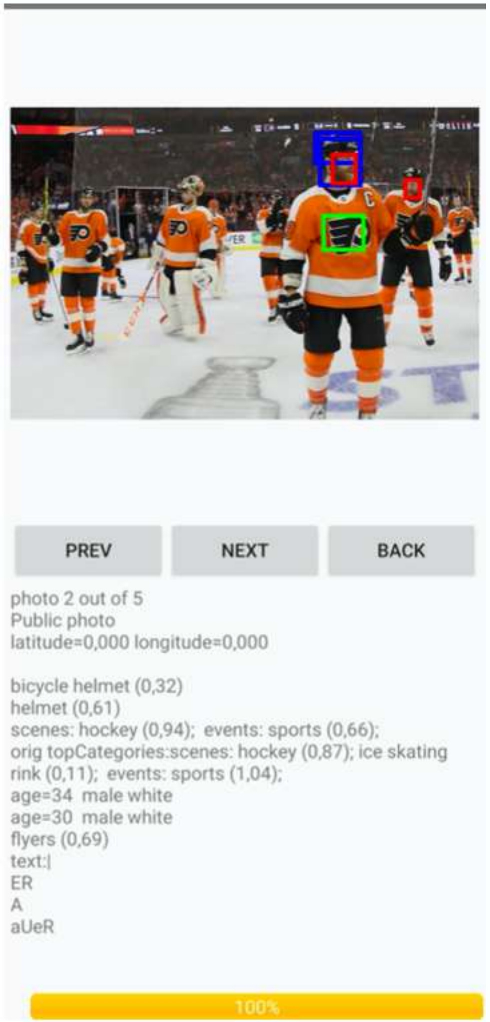
Experimental results (1). Multi-task model for scene recognition, Places2 dataset

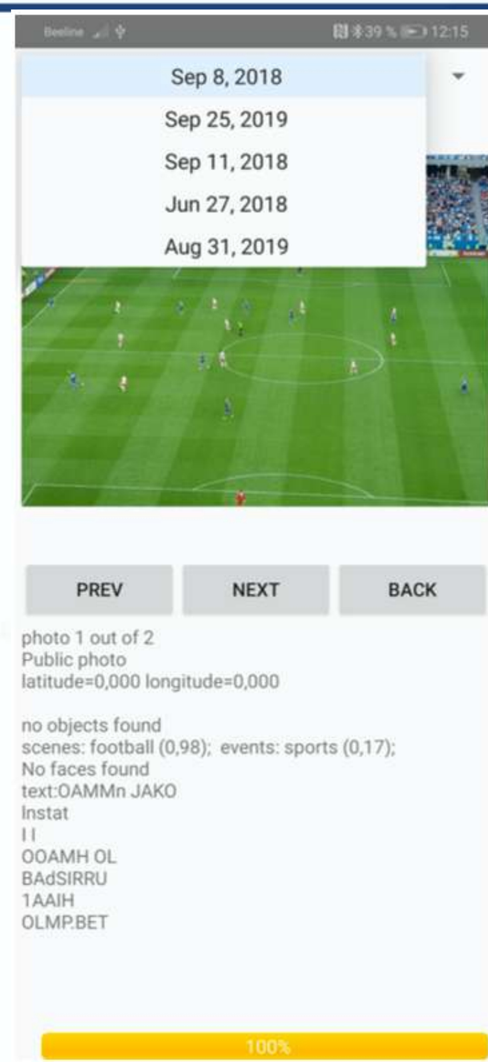
		MobileNet v2 ($\alpha = 1.0$) Original	MobileNet v2 ($\alpha = 1.0$) Structural pruning (25%)	MobileNet v2 ($\alpha = 1.0$) Structural pruning (40%)	MobileNet v2 ($\alpha = 1.4$) Original	MobileNet v2 ($\alpha = 1.4$) Structural pruning (40%)
Size, Mb		11.1	8.3	6.7	20.3	12.2
Inference time, ms	MacBook Pro 2015	18	14	12	31	29
	Galaxy Tab S4	85-105	70-90	60-80	150-180	130-150
	Galaxy S9+	70-90	55-80	50-70	110-160	100-120
All 388 labels	Top-1 accuracy, %	50.7	49.8	48.7	51.3	49.5
	Top-5 accuracy, %	80.4	79.8	79.0	80.7	79.3
	Precision, %	57.5	56.2	54.9	58.0	56.1
	Recall, %	46.7	46.2	45.4	47.1	45.6

Experimental results (2). Multi-task model for event recognition, Photo event collection

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

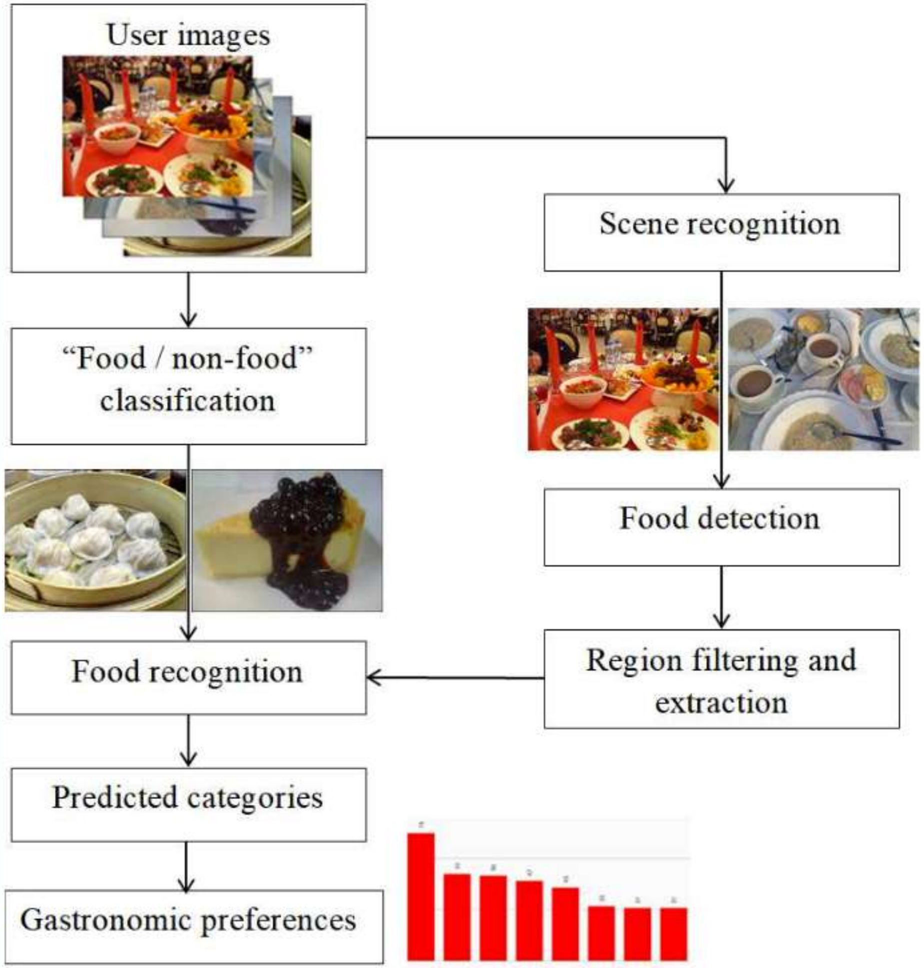
Android demo app

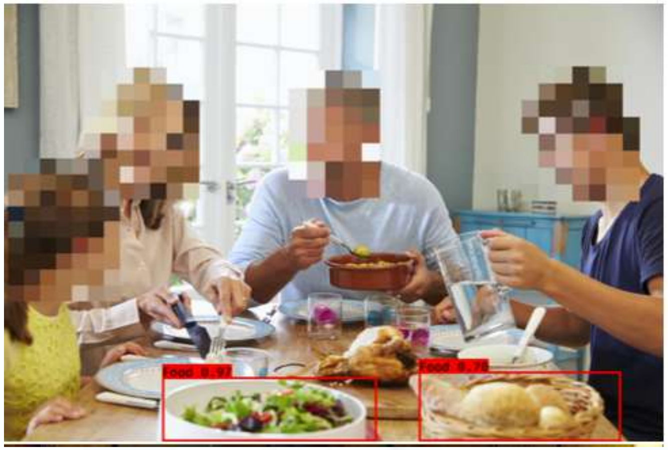




Food classification and restaurant recommendation

Food detection and recognition in a gallery of mobile device





Miasnikov & Savchenko, ICIAR 2020

YELP dataset

Task 1. Restaurant label recognition:
5 classes: drink, food, inside, menu, outside

Task 2. Restaurant Photo Classification Challenge:
Multi-label, 9 classes: good_for_lunch, good_for_dinner, takes_reservations, outdoor_seating, restaurant_is_expensive, has_alcohol, has_table_service, ambience_is_classy, good_for_kids

Task 3. Cuisine recognition:
Multi-label, top-15 classes: American, Italian, …

https://www.yelp.com/dataset/
https://www.kaggle.com/c/yelp-restaurant-photo-classification/data

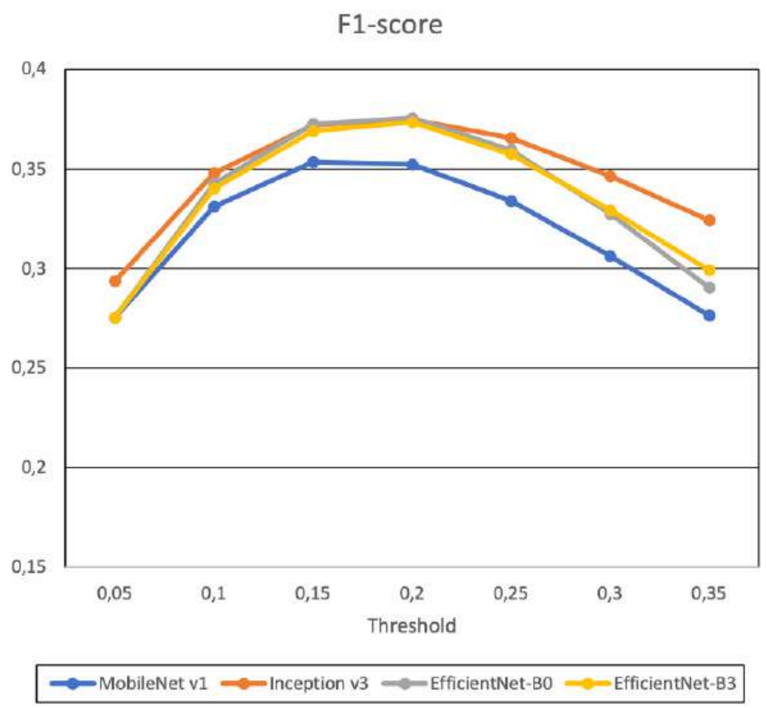
Top-1 accuracy (%) of Yelp restaurant label recognition

ConvNet	Features	Classifier	Top-1 accuracy, %
Mobilenet v2 $\alpha = 1.0$	pre-trained embeddings	Random forest	93.925
	pre-trained embeddings	linear SVM	95.78
	Fine-tuned		97.21
Mobilenet v2 $\alpha = 1.4$	pre-trained embeddings	Random forest	93.545
	pre-trained embeddings	linear SVM	96.19
	Fine-tuned		96.129
Inception v3	pre-trained embeddings	Random forest	94.17
	pre-trained embeddings	linear SVM	96.15
	Fine-tuned		97.15
EfficientNet B5	pre-trained embeddings	Random forest	94.855
	pre-trained embeddings	linear SVM	96.73

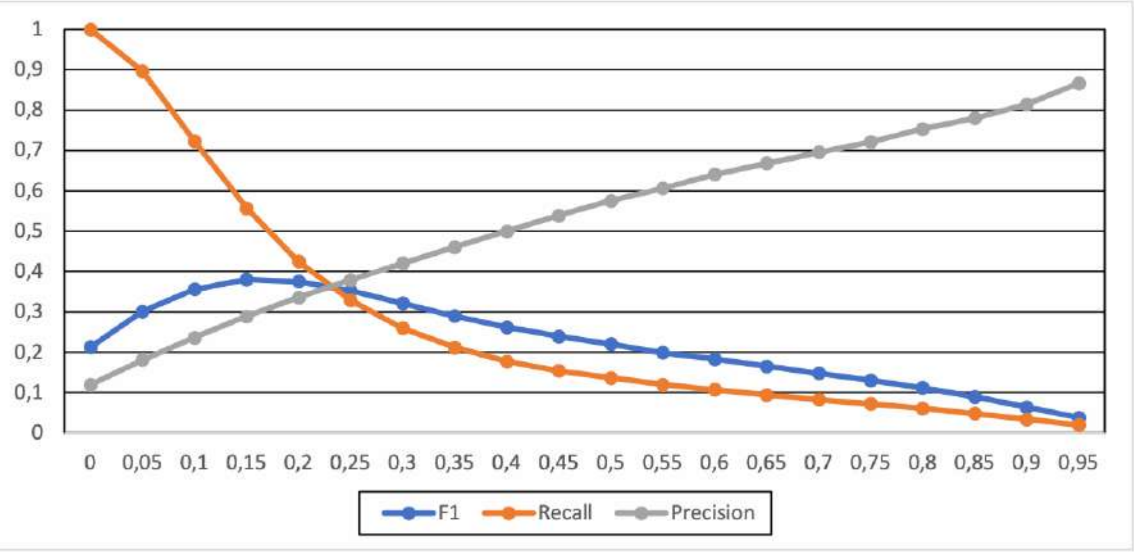
Validation results for the Yelp Restaurant Photo Classification Challenge

Model	Average precision, %	Average recall, %	Average F1 score, %
MobileNet v1	65.39	90.68	75.99
Inception v3	66.91	90.68	77.0

Dependance of F1-score on threshold



Dependance of F1-score/precision/recall on threshold, EfficientNet-B3



Results of cuisine multi-label recognition (threshold=0.15)

	MobileNet v1		Inception v3		EfficientNet-B0		EfficientNet-B3	
	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision
Sandwiches	0.452	0.208	0.456	0.225	0.513	0.227	0.506	0.229
Fast Food	0.436	0.192	0.446	0.238	0.422	0.251	0.412	0.272
American (Traditional)	0.639	0.289	0.609	0.298	0.627	0.293	0.598	0.294
Pizza	0.426	0.238	0.450	0.257	0.419	0.269	0.446	0.249
Breakfast	0.600	0.278	0.590	0.298	0.669	0.275	0.608	0.282
American (New)	0.749	0.385	0.743	0.394	0.813	0.381	0.772	0.393
Italian	0.554	0.241	0.619	0.247	0.593	0.236	0.602	0.244
Mexican	0.273	0.209	0.343	0.238	0.276	0.234	0.297	0.280
Chinese	0.528	0.192	0.557	0.224	0.571	0.217	0.618	0.203
Coffee & Tea	0.305	0.156	0.375	0.191	0.363	0.190	0.396	0.195
Japanese	0.663	0.352	0.659	0.414	0.721	0.373	0.697	0.388
Seafood	0.387	0.298	0.481	0.302	0.464	0.295	0.467	0.298
Sushi Bars	0.509	0.357	0.542	0.406	0.546	0.386	0.542	0.371
Asian Fusion	0.336	0.148	0.382	0.164	0.363	0.148	0.375	0.151
Canadian	0.203	0.137	0.231	0.142	0.154	0.134	0.152	0.133

Examples of food detection and recognition.

Scenes/Events
dining room (0.31) wedding (-0.51)

Objects
Food:0.768
Fast food:0.473
Fireplace:0.422
Fast food:0.407
Snack:0.398
Table:0.389
Snack:0.313
Snack:0.304

YELP label classification
food (0.86)

YELP photo classification
has_alcohol (0.93)

Hotels Classification
Best Western Plus (0.78)

Restaurant recommendation

1. Obtain *C*-dimensional vector **p** of scores (posterior probabilities) of CNN for single image or average scores for all restaurant-related images from a gallery of photos.

2. Multinomial distribution is obtained

$$\tilde{p}_c = \frac{\max(p_c - t_0, 0) + t_0}{\sum_{i=1}^{C} (\max(p_i - t_0, 0) + t_0)}.$$

3. This distribution is used to sample k categories of cuisine.
4. For each cuisine category in this sample, we randomly choose the restaurant from the given city that is associated with this cuisine and has the maximal average number of stars.
5. The set of k restaurants is recommended to a user.

Select city for recommendation Las Vegas

Analyze

Scenes/Events
restaurant (0.62) birthday (-0.06)

Objects
Food:0.692
Furniture:0.604
Bottle:0.602
Plate:0.518
Chopsticks:0.480
Fast food:0.459
Table:0.412
Plate:0.393
Tableware:0.361

YELP restaurants
has alcohol (0.92)
has table service (0.97)
YELP cuisine
Chinese (0.31)
Japanese (0.22)
YELP labels
food (0.99)

Recommended restaurants

Name	Cuisine	Stars
Red Plate	Chinese	5.0
Bar Sake & Robata Grill	Japanese,Sushi Bars	5.0
Ichi Belle	Japanese,Asian Fusion	5.0
Bar Charlie	Japanese	5.0
Tetsuro's Sayonara, Aloha, Going Away Uye At Japanese Curry n	Japanese	5.0

Conclusion

1. Multi-task image recognition improves the decision-making speed.
2. Multi-task learning leads to higher accuracy in many cases, but tuning for one task is still sometimes better
3. Many tasks are still far from maturity:

- sentiments recognition in ads: accuracy 0.37;
- symbolism prediction in ads: F1-score 0.225;
- scene recognition: top-1 accuracy 0.5
- event recognition: accuracy 0.5-0.65;
- cuisine recognition in YELP: F1-score 0.38;
- …

NATIONAL RESEARCH UNIVERSITY

Thank you!