DECISION-MAKING IN VISUAL PRODUCT RECOMMENDATION USING NEURAL AGGREGATION NETWORK AND CONTEXT GATING

- Andrey V. Savchenko
Dr. of Sci., Prof., LATNA, HSE-Nizhny Novgorod
Email: avsavchenko@hse.ru
URL: www.hse.ru/en/staff/avsavchenko
- Kirill Demochkin
NRU HSE-NN, student

Conference on Data-driven Decisions and Analytics (DDA 2018)

NATIONAL RESEARCH UNIVERSITY HIGHER SCHOOL OF ECONOMICS

Outline

1. **Motivation**
2. **Visual product recommendation**
3. **Complete pipeline**
4. **Conclusion and future work**

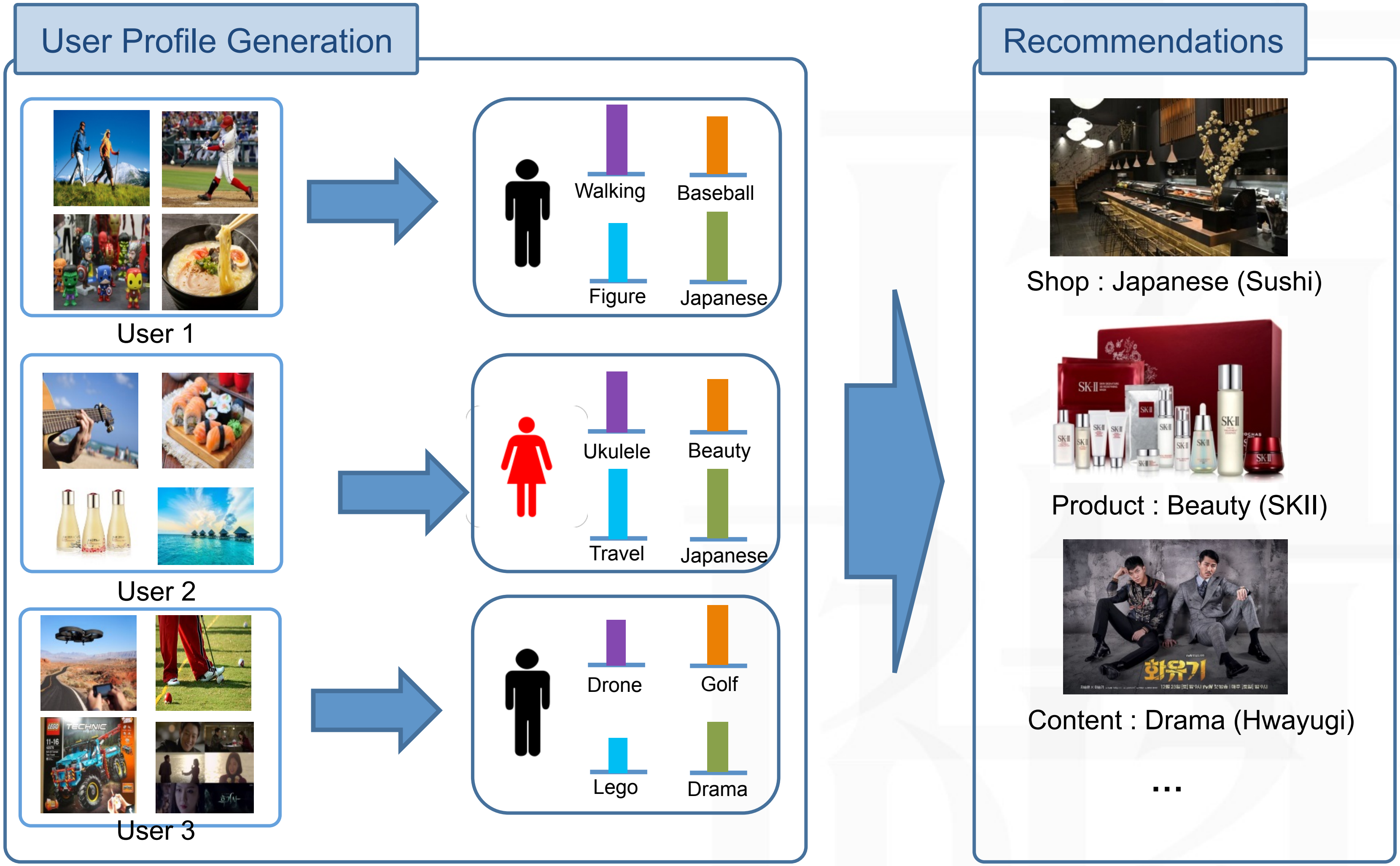
Motivation

User profile

- **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, classification, demographics, hobbies, occupation, lifestyle, etc.) → Generate user profile**

Funded by SAMSUNG

Depending on the user profile, recommends [Product / Shop / Content] to suitable users



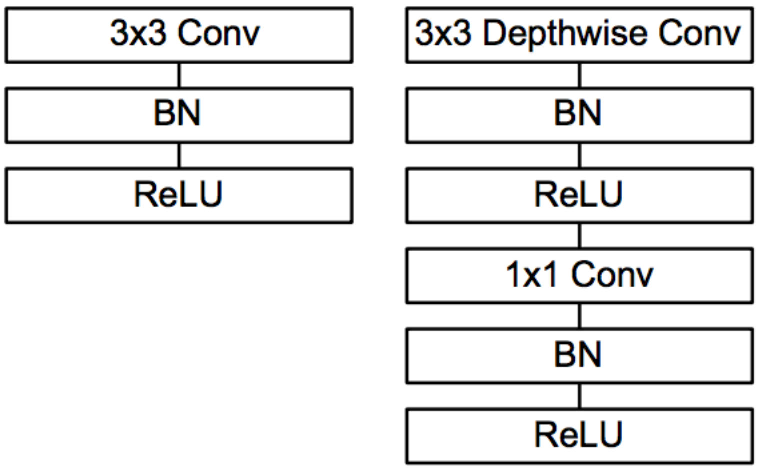
Visual product recommendation

Let there be given N users: the n-th user is associated with M_n images $\{X_n(m)\}$, $m = 1, 2, \ldots, M_n$, of products (single product on each image), that this user has purchased or interacted with. Each product belongs to one or more of D categories.

The task is to predict the relevant classes of products to a user, i.e., use **collection** of images $\{X_n(m)\}$ to generate a D-dimensional vector of scores (estimates of posterior probabilities) that the corresponding category is relevant to the user.

MobileNet

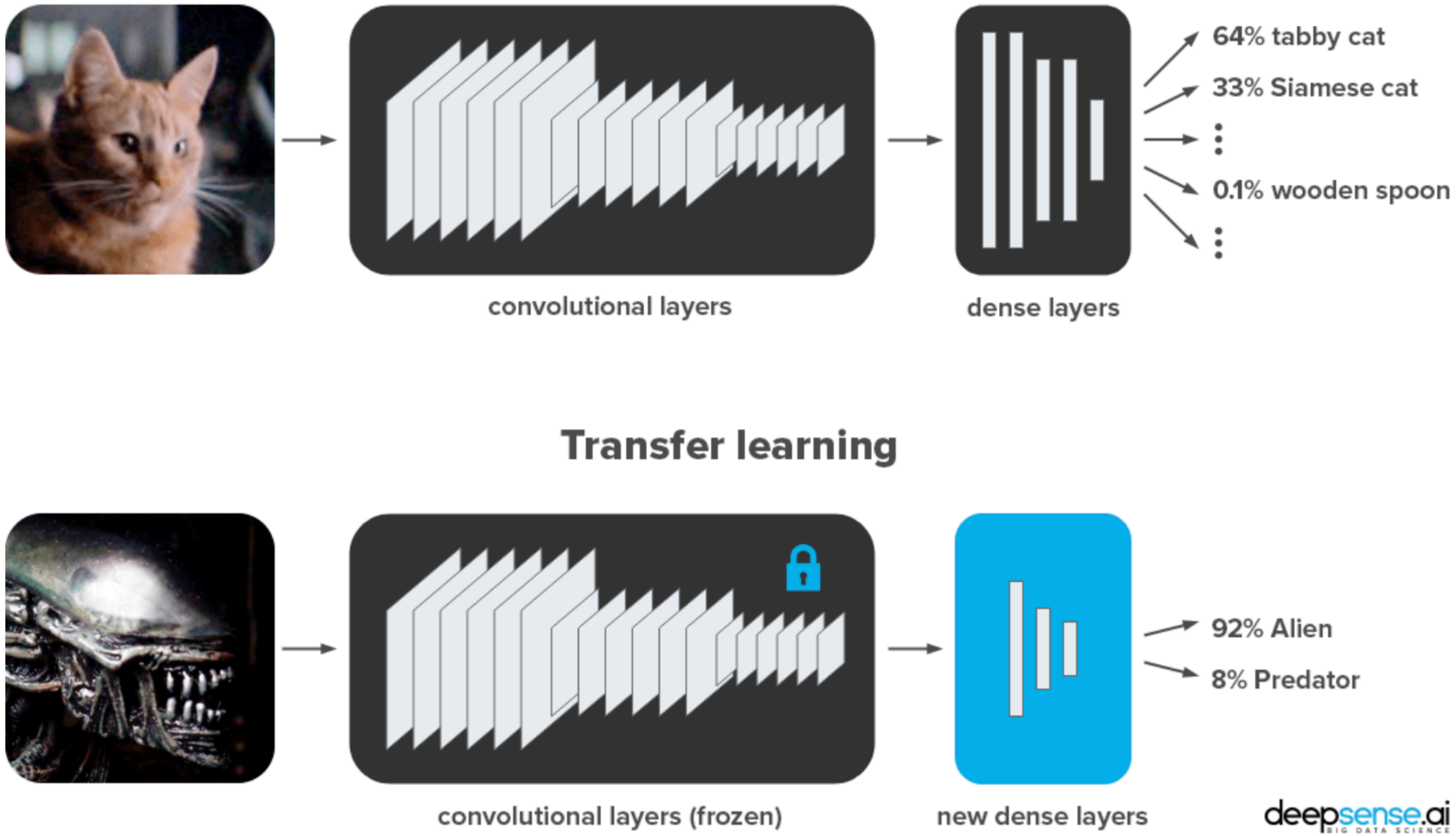
Depthwise convolution



– **28 layers**
– **ImageNet Top-5 Error rate: 12.81%**
– **4.2M parameters**

https://arxiv.org/abs/1704.04861

Type / Stride	Filter Shape	Input Size
Conv / s2	$3 \times 3 \times 3 \times 32$	$224 \times 224 \times 3$
Conv dw / s1	$3 \times 3 \times 32$ dw	$112 \times 112 \times 32$
Conv / s1	$1 \times 1 \times 32 \times 64$	$112 \times 112 \times 32$
Conv dw / s2	$3 \times 3 \times 64$ dw	$112 \times 112 \times 64$
Conv / s1	$1 \times 1 \times 64 \times 128$	$56 \times 56 \times 64$
Conv dw / s1	$3 \times 3 \times 128$ dw	$56 \times 56 \times 128$
Conv / s1	$1 \times 1 \times 128 \times 128$	$56 \times 56 \times 128$
Conv dw / s2	$3 \times 3 \times 128$ dw	$56 \times 56 \times 128$
Conv / s1	$1 \times 1 \times 128 \times 256$	$28 \times 28 \times 128$
Conv dw / s1	$3 \times 3 \times 256$ dw	$28 \times 28 \times 256$
Conv / s1	$1 \times 1 \times 256 \times 256$	$28 \times 28 \times 256$
Conv dw / s2	$3 \times 3 \times 256$ dw	$28 \times 28 \times 256$
Conv / s1	$1 \times 1 \times 256 \times 512$	$14 \times 14 \times 256$
$5\times$ Conv dw / s1	$3 \times 3 \times 512$ dw	$14 \times 14 \times 512$
$5\times$ Conv / s1	$1 \times 1 \times 512 \times 512$	$14 \times 14 \times 512$
Conv dw / s2	$3 \times 3 \times 512$ dw	$14 \times 14 \times 512$
Conv / s1	$1 \times 1 \times 512 \times 1024$	$7 \times 7 \times 512$
Conv dw / s2	$3 \times 3 \times 1024$ dw	$7 \times 7 \times 1024$
Conv / s1	$1 \times 1 \times 1024 \times 1024$	$7 \times 7 \times 1024$
Avg Pool / s1	Pool 7×7	$7 \times 7 \times 1024$
FC / s1	1024×1000	$1 \times 1 \times 1024$
Softmax / s1	Classifier	$1 \times 1 \times 1000$



https://deepsense.ai/keras-vs-pytorch-avp-transfer-learning/

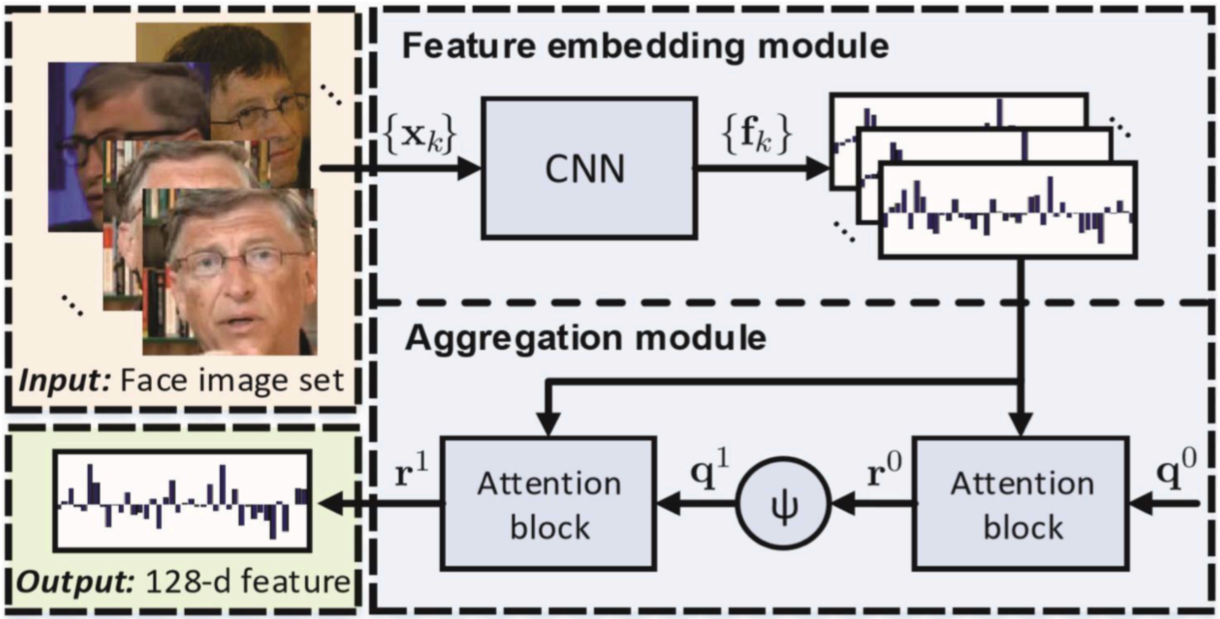
How to aggregate visual features from ConvNet for several images associated with a user?

(Weighted) average

$$x_n = \sum_{m=1}^{M_n} w(x_n(m))x_n(m)$$

1. Simple average (all weights are equal to $1/M_n$)
2. Learnable pooling:

- Neural aggregation network
- Context gating



Average weighting

$$\mathbf{r} = \sum_k a_k \mathbf{f}_k$$

Attention block

$$e_k = \mathbf{q}^T \mathbf{f}_k$$

$$a_k = \frac{\exp(e_k)}{\sum_j \exp(e_j)}.$$

Attention (2nd block)

$$\mathbf{q}^1 = \tanh(\mathbf{W}\mathbf{r}^0 + \mathbf{b})$$

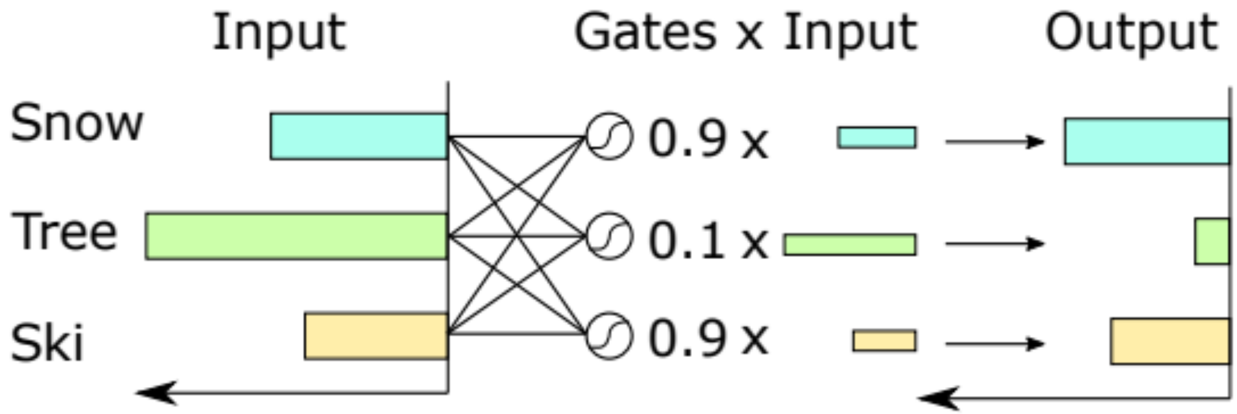
CVPR17 (https://arxiv.org/abs/1603.05474)

The Context Gating (CG) module transforms the input feature representation X into a new representation Y as

$$Y = \sigma(WX + b) \circ X$$

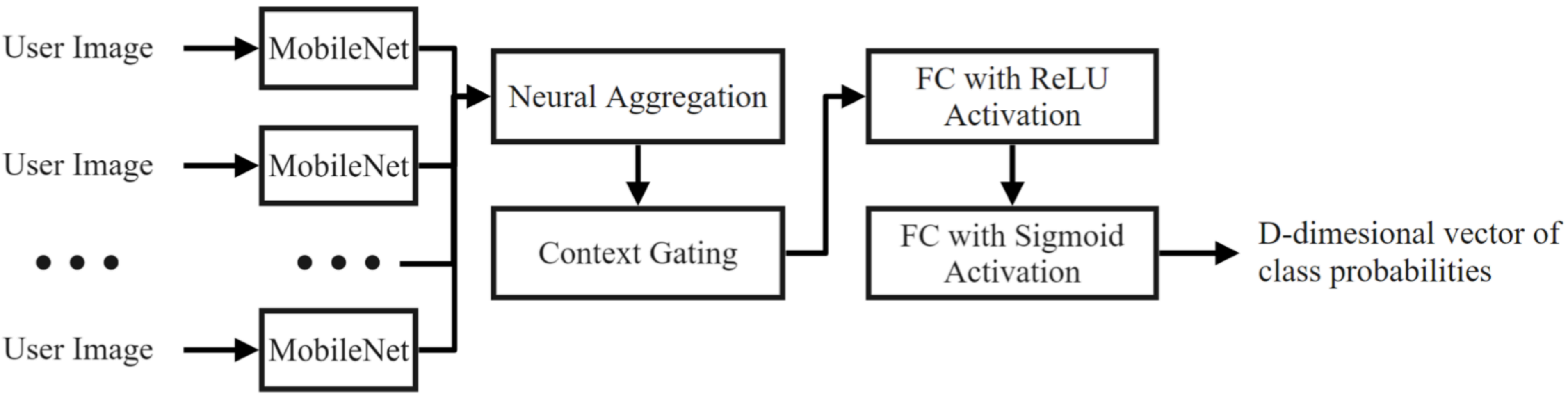
$$\nabla Y = \nabla(\sigma(WX + b)) \circ X + \sigma(WX + b) \circ \nabla X$$

CG down-weights visual activations of Tree for a skiing scene

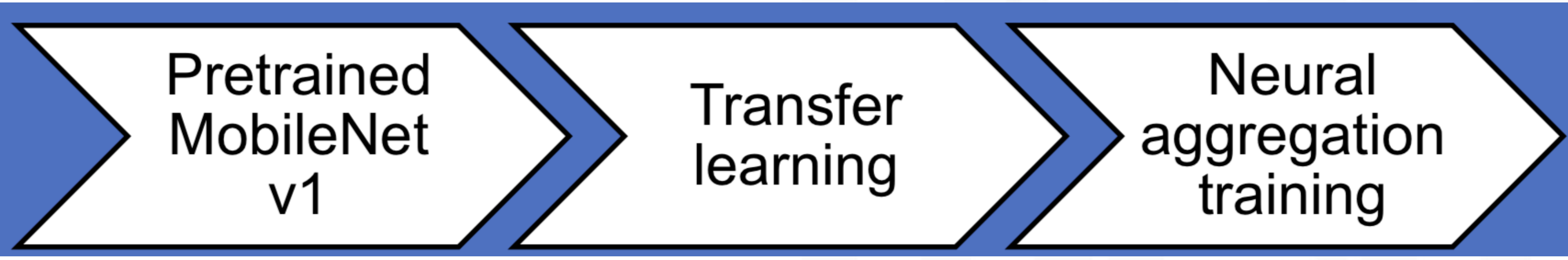


Youtube8M CVPR17 workshop (https://arxiv.org/abs/1706.06905)

Proposed network



Training pipeline



Experiments. Amazon Products Dataset

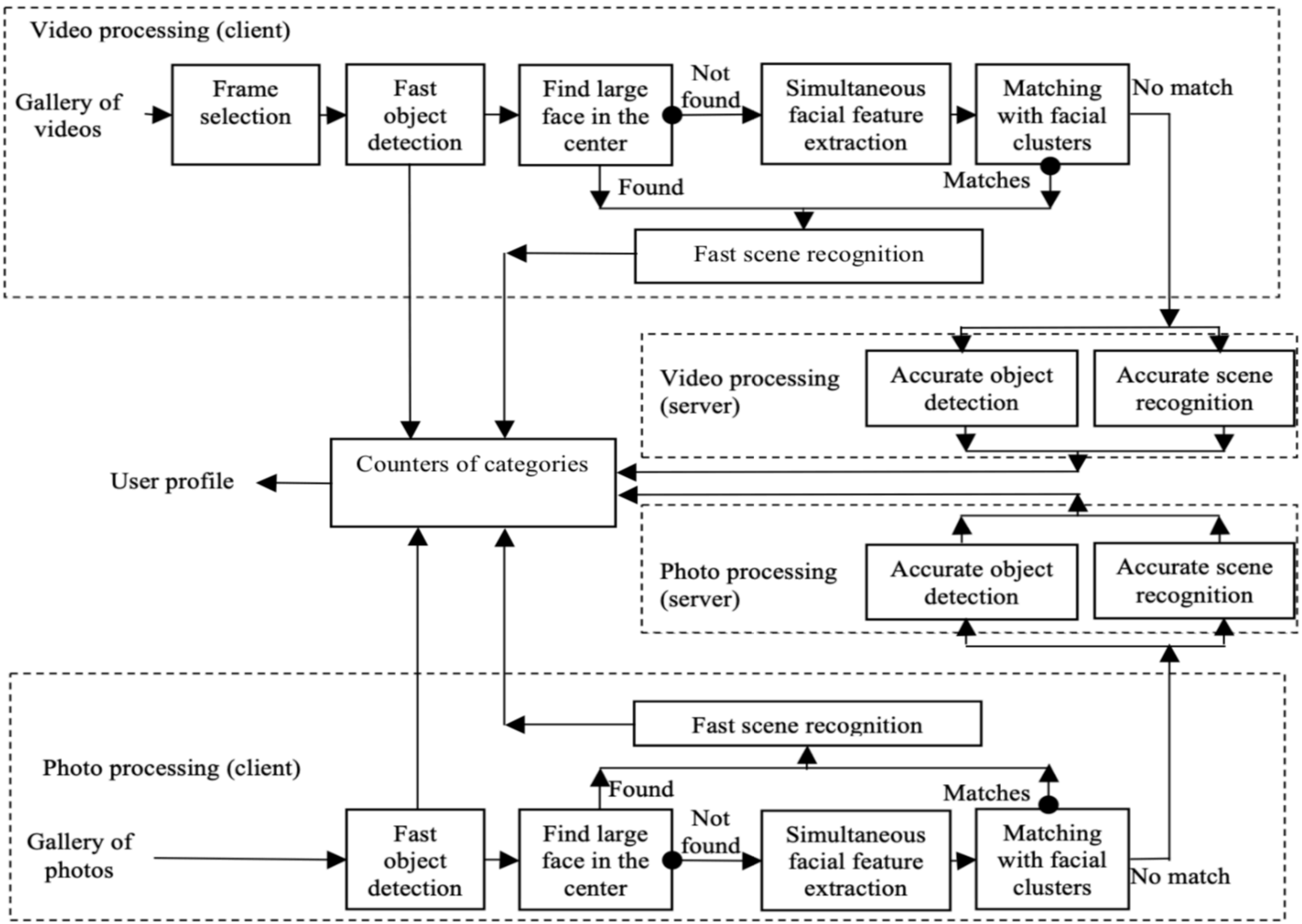
547700 entries
66519 unique users
28237 unique items
1000 categories

Experiments. Recall@k and Precision@k for different aggregation strategies

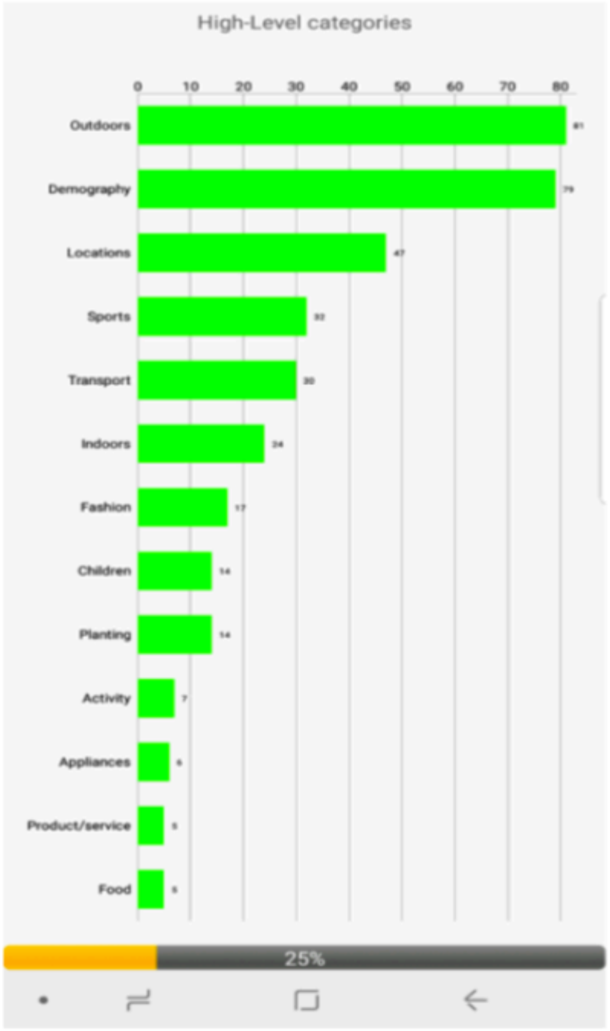
k	Aggregation	Recall @k	Precision @k
5	Average	0.704867	0.749925
	Neural Aggregation	0.772574	0.839458
	Neural Aggregation + Context Gating	**0.792203**	**0.922438**
10	Average	0.797340	0.595867
	Neural Aggregation	0.901716	0.710123
	Neural Aggregation + Context Gating	**0.91846**	**0.881151**
15	Average	0.815469	0.561431
	Neural Aggregation	0.932418	0.710123
	Neural Aggregation + Context Gating	**0.942565**	**0.868210**
20	Average	0.820141	0.553453
	Neural Aggregation	0.943513	0.636783
	Neural Aggregation + Context Gating	**0.947498**	**0.864384**

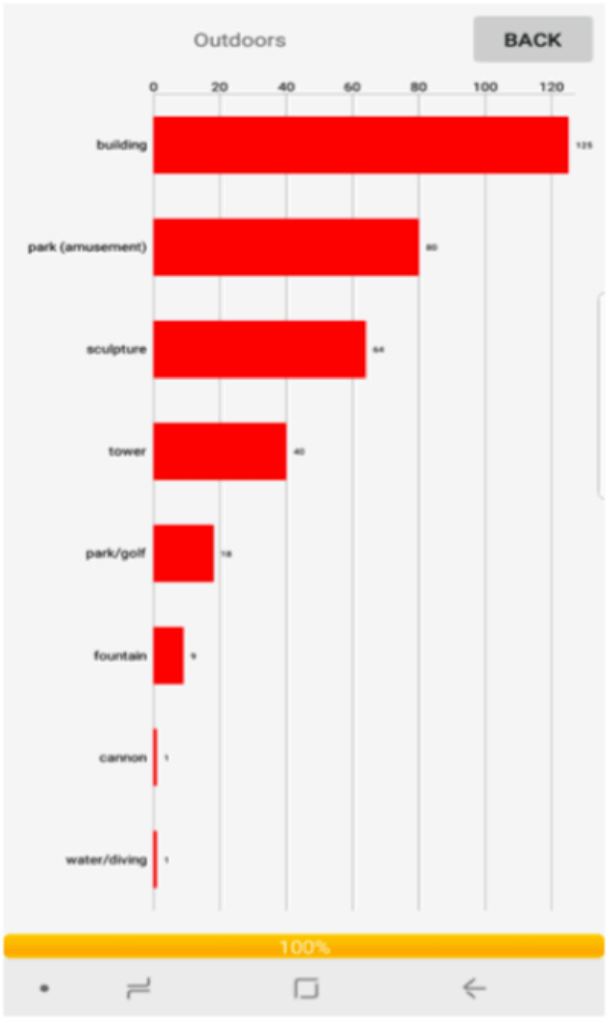
Complete pipeline

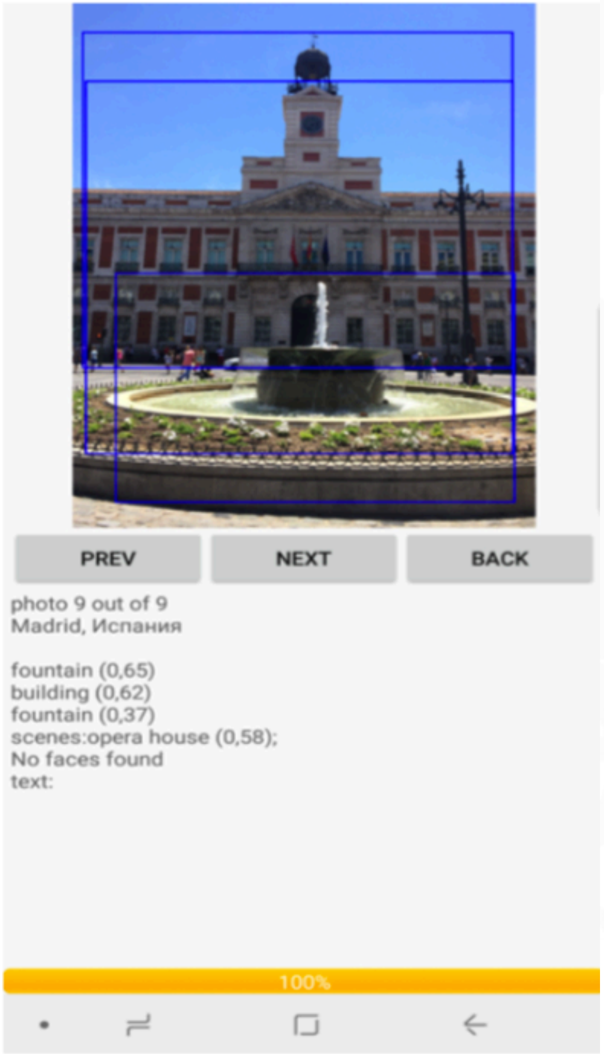
Proposed pipeline for visual preferences prediction



Example


High-Level categories
0 10 20 30 40 50 60 70 80
Outdoors
Demography
Locations
Sports
Transport
Indoors
Fashion
Children
Planting
Activity
Appliances
Product/service
Food
25%


Outdoors
BACK
0 20 40 60 80 100 120
building
park (amusement)
sculpture
tower
park/golf
fountain
cannon
water/diving
100%


PREV
NEXT
BACK
photo 9 out of 9
Madrid, Испания
fountain (0,65)
building (0,62)
fountain (0,37)
scenes:opera house (0,58);
No faces found
text:
100%

Conclusion

1. The neural aggregation with context gating outperforms the naive averaging method by up to 34%
2. We obtained the state-of-the-art results in video-based age prediction and gender recognition based on special multi-output MobileNet, the Dempster-Shafer theory for aggregation of predicted gender posterior probabilities and computation of mean expectation by using the top-k predicted ages
3. We considered the scene recognition task with 350 different scenes and obtained 89% top-5 accuracy using the most powerful pipeline
4. We implemented the complete pipeline for organizing photo and video albums based on facial clustering and obtained the state-of-the-art results for the recently appeared GFW dataset
5. We prepared three Russian patent applications in cooperation with Samsung R&D Institute Russia

Future work

1. **Develop mobile recommender system**
2. **Offline conversion of images to text descriptions (image captioning) and analysis of resulted textual data;**
3. **Recognition of specific items (types of food, pet breed, etc.)**
4. **Include face re-identification into the current app in order to analyze the demography more accurately**
5. **Extract user preferences from the text recognized in images**

HIGHER SCHOOL OF ECONOMICS

NATIONAL RESEARCH UNIVERSITY

Thank you!