Visual product recommendation using neural aggregation network and context gating

Visual product recommendation using neural aggregation network and context gating

K V Demochkin[1], A V Savchenko[1]

[1]National Research University Higher School of Economics, Rodionov street, 136, Nizhny Novgorod, Russia

e-mail: kvdyomochkin@edu.hse.ru, avsavchenko@hse.ru

Abstract. In this paper we focus on the problem of user interests' classification in visual product recommender systems. We propose the two-stage procedure. At first, the visual features are learned by fine-tuning the convolutional neural network, e.g., MobileNet. At the second stage, we use such learnable pooling techniques as neural aggregation network and context gating in order to compute a weighted average of image features. As a result we can capture the relationships between the products images purchased by the same user. We provide an experimental study with the Amazon product dataset. It was shown that our approach achieves a F1-score of 0.90 for 15 recommendations, which is much higher when compared to 0.66 F1-measure classification of traditional averaging of the feature vector.

1. Introduction

There is a recent rise in interest concerning visual recommender systems [1, 2, 3, 4] that infer user preferences (predict categories of interest for a user) by analysing a set of pictures of items that the user either bought or browsed earlier. Such systems may be used independently or as the core of existing recommender systems in online shops for fast and reliable estimation of categories of products that might be of interest to a particular user based on the information gathered from, e.g., a mobile application. In the recent years several visual feature extraction methods have been developed that reinforce the idea that specific approaches are required in order to learn robust visual features [5, 6, 7].

The categories of products that a user is interested in are often correlated. Hence, in this paper we propose to use modern learnable pooling techniques to capture the interdependencies between images of the same user. Such methods were originally developed for video recognition tasks, for example, face identification and verification on video [8, 9, 10]. Among such techniques the most successful are the neural aggregation network [11], and the context gating [12] that won the prestigious Youtube 8M Large-Scale Video Understanding challenge 2017.

Therefore, the goal of this paper is to develop a user modelling engine for visual recommender systems based on combination of known techniques of weighted aggregation of image features, previously used for video analysis tasks. The reported results and conclusions are aimed at a wide range of experts in computer vision and recommender systems.

2. Literature survey

Traditional video recognition methods [8] usually represent each frame as a high-dimensional feature vector, extracted from one of the last layers of a deep convolutional neural network. Next, an ensemble of classifiers is possibly applied to make one decision based on all video frames [13]. For example, authors of article [14] modified the probabilistic approach to face recognition to work with a collection of images and video streams. In the paper [15] it is suggested that the feature vectors for all video frames are uniformly distributed and the Kullback-Leibler divergence can be used to measure the distance between distributions. Several important works on video classification are based on metric learning [16]. For instance, a metric for estimating the similarity between an image and a collection of images and the similarity between two collections of images has been successfully trained in [17].

To speed up the decision-making process, the feature vectors of each image may be combined into a single vector with average or max pooling [18, 19]. Moreover, aggregation methods with trainable weights ("learnable pooling") are receiving all the more attention. One such technique, called Eigen-PEP [19] embedding integrates visual information from all images by using partial averaging based on a probabilistic model of elastic parts. Afterwards, the intermediate representation is compressed using the primary component analysis method. Canziani and Culurciello presented the CortexNet [20] to extract robust and stable representations of time changing signals. In [21] two video streams with head movements and various facial expressions were compared with a positive definite kernel based on calculation of angles between two linear subspaces. Miech et al. [12] won the prestigious YouTube 8M Large-Scale Video Understanding challenge in 2017 with an approach based on learnable pooling methods such as Soft Bag-of-words, Fisher Vectors, NetVLAD [22, 23] and context gating to model interdependencies between different classes. Interesting results were reported by Yang et al. [11] who proposed to train dynamic weights for frames in a video stream with a neural aggregation system that consists of two sequential attention blocks.

Consequently, methods for trainable aggregation of feature vectors were originally developed as promising solutions to the video classification task and simple object classification. In this paper we decided to apply the most prominent approaches to classification of *a set of images* in order to create a recommender system based on effective and efficient classification algorithms.

3. Proposed approach

The image-based user interest prediction task can be formulated as follows: it is required to predict the relevant classes of products to a user based on a collection of images of products that this user has previously bought. Every product belongs to one or more of D categories. In other words, the goal is to estimate posterior probabilities that a user orders a product from each of the D categories. It is supposed that a collection $\{X_n(m)\}$, m=1,2, …, M_n of M_n images of products that this user has purchased or browsed is available to train a classifier. We assume that there is a single unique item on each picture that belongs to one or more of D categories, so that each image is associated with a binary label vector $\mathbf{y}$ of length D, which d-th component is set to 1 if the item on the image belongs to the i-th category and 0 otherwise.

We propose the following two stage approach for such multi-label image set classification task. At first, transfer learning [24] is applied for feature extraction by adding a classifier to a base deep convolution neural network that was pre-trained on a large set of images such as ImageNet-1000 [25]. Considering the constraints imposed on the system for running on a mobile device, which were discussed in the introduction, the MobileNet [26] architecture was chosen for this paper. We split the N collections of images into two disjoint sets with size N_1 and N_2. The first set is used in the fine-tuning process to learn the feature extractor. The fine-tuned model is then used to obtain the K-dimensional feature vectors $\mathbf{x}_n(m)$ for each image in the second subset.

Secondly, these features are aggregated into single K-dimensional descriptor $\mathbf{x}_n$ of the n-th user by computing a weighted sum of features of individual images:

$$\mathbf{x}_n = \sum_{m=1}^{M_n} w(\mathbf{x}_n(m))\mathbf{x}_n(m), \tag{1}$$

where the weights may depend on the features $\mathbf{x}_n(m)$. If the equal weights are used then conventional averaging with computation of mean feature vector is implemented [11]. However, in this paper we analyze the neural network-based methods with learning of weights in (1), particularly, the neural aggregation module with an attention mechanism originally used in video-based face recognition [11]:

$$w(\mathbf{x}_n(m)) = \frac{\exp(\mathbf{q}\mathbf{x}_n(m))}{\sum_{j=1}^{M_n} \exp(\mathbf{q}\mathbf{x}_n(j))}. \quad (2)$$

Here $\mathbf{q}$ is the K-dimensional vector of learned parameters. Moreover, we additionally use the context gating [12]:

$$\mathbf{x}_n^{(1)} = \sigma(W\mathbf{x}_n + b) \circ \mathbf{x}_n, \quad (3)$$

where σ() is logistic sigmoid function, symbol ∘ stands for element-wise multiplication, and matrix W and bias b are the learned weights of this layer. Context gating (3) applies a scaling mask to the resulting aggregated vector (3) for modelling the interdependencies of different categories and estimating the categories that often appear together. Hence, the weights for closely related categories should be scaled up if they are present in a single collection. The opposite is also true: for categories that are not likely to appear simultaneously the weights are reduced.

The complete model architecture is shown in Figure 1. Here the aggregated vectors (3) are passed to a fully connected layer with dropout regularization. Because the resulting vector $\mathbf{y}$ of labels often contains multiple non-zero elements since the product may belong in more than a single category, the output layer has sigmoid activation. Consequently, the output layer predicts the posterior probabilities that the d-th category is relevant to the particular user.

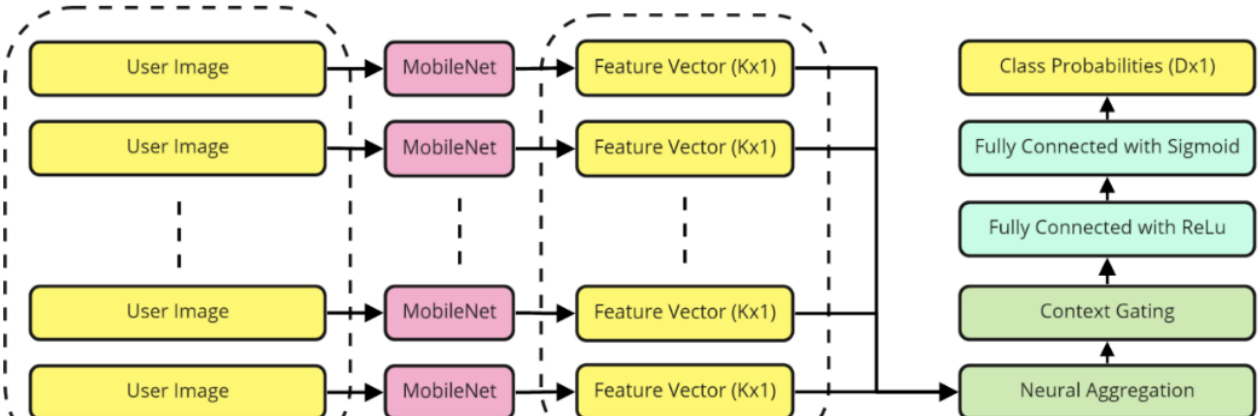

Figure 1. Proposed neural network architecture for item recommendation based on an image collection.

4. Experimental Results

In the experimental study the "Home and Kitchen" 5-core subset of the Amazon Product Data dataset [27] was used (Figure 2), meaning only the items that have at least 5 unique users interacted with and only users that have interacted with at least 5 unique items are kept. Such subset contains 547700 entries of N=66519 unique users interacting with 28237 unique items from D=1000 categories. The list of categories includes "Cookware", "Storage & Organization", "Coffee", etc. For each user there is data available on the items that the user has bought. The number of items per user M_n varies from 5 to 40; the average user has interacted with 8 unique items. Each user was assigned a D-dimensional vector y where d-th element is 0 if the user has not bought any items from category d, and 1 otherwise if the user has bought at least a single product from category ***d***. All experiments were conducted on a single Nvidia GeForce GTX 1080ti GPU. The algorithms were implemented using the Keras framework.

As part of the preprocessing stage each image was resized to 224x224 pixels and RGB values were normalized to the range [-1;1] to conform to the input format required by the MobileNet v1 pre-trained

with ImageNet weights. At first, 70% of all images (N_1=0.7N=46563) selected by random split were used to fine-tune the MobileNet model.

Figure 2. Examples of images from the "Home and Kitchen" subset.

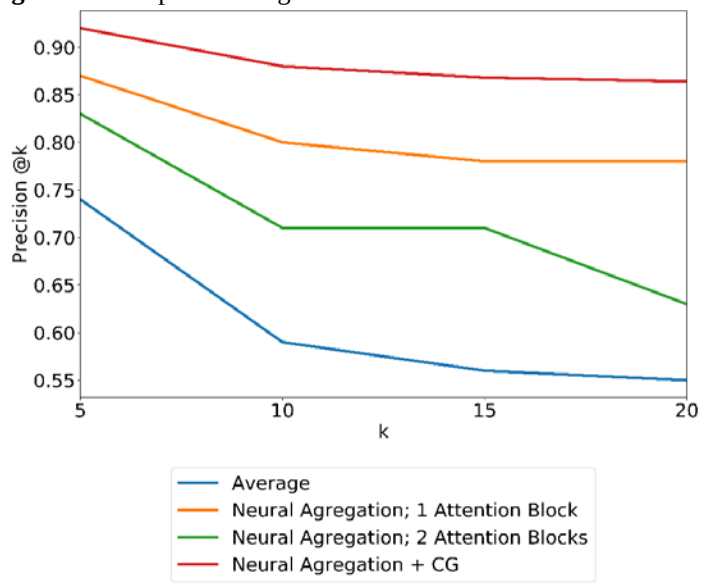

Figure 3. Dependency of precision on different number *k* of recommendations.

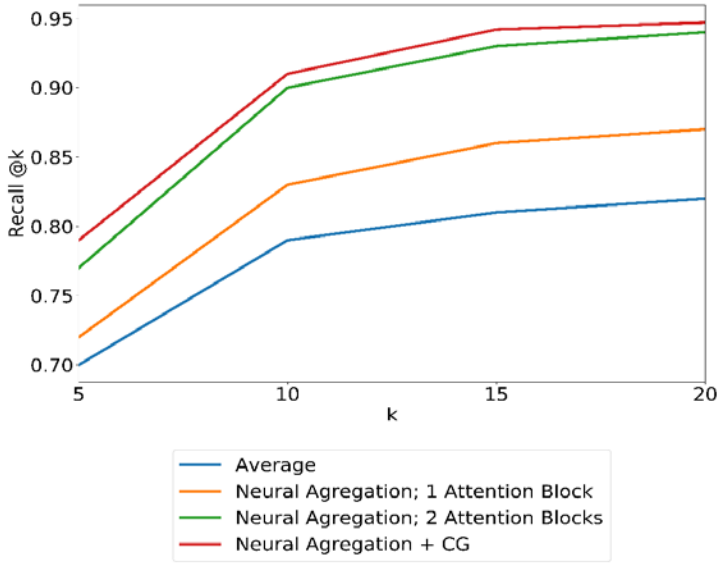

Figure 4. Dependency of recall on different number *k* of recommendations.

It should be noted that since each item belongs to only a few of the categories the resulting target vectors are sparse. To reduce the class imbalance, we implemented the weighted binary cross-entropy objective function. Several values were tested for the positive class weight {10, 36, 72, 140} and the best quality was achieved for positive class weight equal to 36. We also found that the hidden layer with 2048 neurons and 50% dropout probability worked best in our experiment. The model was first trained with 22 frozen deepest layers in batches of 64 samples using the ADAM optimizer with a learning rate of 0.001, $\beta_1 = 0.9$, $\beta_2 = 0.999$ for 10 epochs. Then it was trained with all layers unfrozen for 20 more epochs with only the learning rate changed to 0.0001.

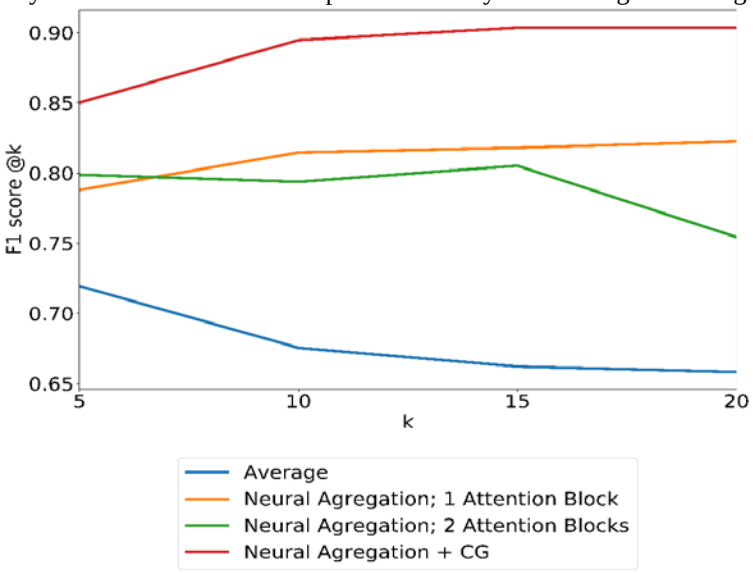

Figure 5. Dependency of the F1 score on different number *k* of recommendations.

As for the aggregation step, we tested four approaches (Average pooling, Neural Aggregation with a single attention block, Neural Aggregation with two sequential attention blocks, Neural Aggregation + Context Gating). For the first approach an average of the feature vectors was obtained. For the second and third approaches one and two attention blocks were utilized respectively similar to [11]. For the proposed approach the two sequential attention blocks were followed by a context gating layer [12], which dynamically rescales the aggregated feature vector using trainable weights.

After the features are pooled into a single vector, it is passed to one dense hidden layer of size 2048 with ReLU activation function, which output is fed into a linear layer with sigmoid activations that predicts the final relevance of categories. 70% of the second subset was used to learn the pooling weights while the remaining 30% of user data was used for testing. The weighted cross entropy function with positive weight equal to 36 was used as the loss. The model was trained with the ADAM optimizer with learning rate = 0.001, beta_1 = 0.9, beta_2 = 0.999.

The dependence of the F1-measure on the number of recommendations k as well as details about precision @k and recall @k [28] are presented in Figures 3, 4, and 5. Here the highest F1-measure is achieved by the combination of neural aggregation [11] and context gating [12] is 12-35% higher when compared to simple averaging of visual features. The addition of Context Gating to the neural aggregation network makes it possible to improve the decision-making quality in 5-14%. It should be noted that for averaging the precision falls off sharply by almost 0.2 as the number of top recommendations is increased. Moreover, recall growth is most apparent from 5 to 10 recommendations, however further increasing the parameter *k* does not have much of an effect, as it levels out at around 0.8. The most stable precision is obtained via our proposed method as it only degrades by 0.06 from 0.92 at *k*=5 to 0.86 at *k*=20. Interestingly, a single neural aggregation block has consistently higher precision than two consequent neural aggregation blocks [11], and despite the recall metric being in favor of two aggregation blocks [11] the later has a lower F1 score starting at *k*=10. This observation could be attributed to the fact that two neural aggregation blocks [11] need a smaller positive weight for the weighted cross entropy loss function in order to compensate for their increased capacity, which leads to overfitting to positive samples and, consequently, high recall and low precision.

5. Conclusion and Future Work

This paper demonstrates the application of video data analysis methods of aggregation to the task of user interest prediction based on a collection of images of products that the user has previously shopped for. It was experimentally shown that the neural aggregation [11] with context gating [12] outperforms the simple averaging method by up to 34% (Figure 5).

The main direction for further research is to expand the proposed approach into a complete mobile recommender system that would suggest the items from relevant categories to that user based only on the images on the user's device. Additionally, it is imperative to compare the performance of our approach with traditional recommender system methods, e.g. collaborative filtering or factorization machines [29]. Finally, it is vital to work with other open datasets, e.g., the Amazon Fashion dataset that features collections of clothing items that were bought by Amazon users.

6. References

[1] Hidasi B, Quadrana M, Karatzoglou A and Tikk D 2016 Parallel Recurrent Neural Network Architectures for Feature-rich Session-based Recommendations *Proc. of the 10th ACM Conf. on Recommender Systems* (New York: ACM) 241-248

[2] Shankar D, Narumanchi S, Ananya H A, Kompalli P and Chaudhury K 2017 Deep learning based large scale visual recommendation and search for e-commerce *Preprint* arXiv:1703.02344

[3] Andreeva E, Ignatov D I, Grachev A and Savchenko A 2018 Extraction of Visual Features for Recommendation of Products via Deep Learning *Proc. of Int. Conf. on Analysis on Images, Social Networks and Texts – AIST* (New York: Springer) 201-210

[4] Zhai A, Kislyuk D, Jing Y, Feng M, Tzeng E, Donahue J and Darrell T 2017 Visual discovery at pinterest *Proc. of the 26th Int. Conf. on World Wide Web Companion* 515-524

[5] Myasnikov E V 2017 Hyperspectral image segmentation using dimensionality reduction and classical segmentation approaches *Computer Optics* **41(4)** 564-572 DOI: 10.18287/2412-6179-2017-41-4-564-572

[6] Savchenko A V 2018 Trigonometric series in orthogonal expansions for density estimates of deep image features *Computer Optics* **42(1)** 149-158 DOI: 10.18287/2412-6179-2018-42-1-149-158

[7] Savchenko A V 2017 Maximum-likelihood dissimilarities in image recognition with deep neural networks *Computer Optics* **41** 422-430 DOI: 10.18287/2412-6179-2017-41-3-422-430

[8] Sokolova A D and Savchenko A V 2018 Data organization in video surveillance systems using deep learning *CEUR Workshop Proceedings* **2210** 243-250

[9] Nikitin M Y, Konouchine V S and Konouchine A S 2017 Neural network model for video-based face recognition with frames quality assessment *Computer Optics* **41** pp 732-742 DOI: 10.18287/2412-6179-2017-41-5-732-742

[10] Sokolova A D and Savchenko A V 2018 Cluster analysis of facial data in video surveillance systems using deep learning *Computational Aspects and Applications in Large-Scale Networks* **7** 113-120

[11] Yang J, Ren P, Zhang D, Chen D, Wen F, Li H and Hua G 2017 Neural aggregation network for video face recognition *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* **4** 7

[12] Miech A, Laptev I and Sivic J 2017 Learnable pooling with context gating for video classification *Preprint* arXiv:1706.06905

[13] Valentini G and Masulli F 2002 Ensembles of learning machines *Proc. Italian Workshop on Neural Nets* (New York: Springer New York LLC) 3-20

[14] Zhang Y and Martinez A M 2006 A weighted probabilistic approach to face recognition from multiple images and video sequences *Image and Vision Computing* **24** 626-638

[15] Shakhnarovich G, Fisher J W and Darell T 2002 Face recognition from long-term observations *Proc. European Conference on Computer Vision* (New York: Springer New York LLC) 851-865

[16] Huang Z, Wang R, Shan S and Chen X 2014 Learning Euclidean-to-Riemannian metric for

point-to-set classification *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition* (Washington: IEEE Computer Society) 1677-1684
[17] Zhu P, Zhang L, Zuo W and Zhang D 2013 From point to set: Extend the learning of distance metrics *Proc. of the IEEE International Conference on Computer Vision* (Washington: IEEE Computer Society) 2664-2671
[18] Chen J C, Ranjan R, Kumar A, Chen C H, Patel V M and Chellappa R 2015 An end-to-end system for unconstrained face verification with deep convolutional neural networks *Proc. of the IEEE International Conference on Computer Vision Workshops* (Washington: IEEE Computer Society) 118-126
[19] Li H, Hua G, Shen X, Lin Z and Brandt J 2014 Eigen-pep for video face recognition *Proc. Asian Conference on Computer Vision* 17-33
[20] Canziani A and Culurciello E 2017 Cortexnet: a generic network family for robust visual temporal representations *Preprint* arXiv:1706.02735
[21] Wof L and Shashua A 2003 Kernel principal angles for classification machines with applications to image sequence interpretation *Proc. of the IEEE Computer Society Conference* **1** 1
[22] Arandjelovic R, Gronat P, Torii A, Padjla T and Sivic J 2016 Netvlad: Cnn architecture for weakly supervised place recognition *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition* 5297-5307
[23] Perronnin F and Dance C 2007 Fisher kernels on visual vocabularies for image categorization *Proc. IEEE Conference on Computer Vision and Pattern Recognition* (Washington: IEEE Computer Society) 1-8
[24] Pan S J and Qiang Y 2010 A survey on transfer learning *IEEE Transactions on knowledge and data engineering* **22(10)** 1345-1359
[25] Krizhevsky A, Sutskever I and Hinton G E 2010 ImageNet classification with deep convolutional neural networks *Advances in neural information processing systems* 1097-1105
[26] Howard A G, Zhu M, Chen B, Kalenichenko D, Wang W, Weyand T and Adam H 2017 MobileNets: efficient convolutional neural networks for mobile vision applications *Preprint* arXiv:1704.04861
[27] McAuley J, Targett C, Shi Q and Van Den Hengel A 2015 Image-based recommendations on styles and substitutes *Proc. of the 38th International ACM SIGIR Conf. on Research and Development in Information Retrieval* 43-52
[28] Herlocker J L, Konstan J A, Terveen L G and Riedl J T 2004 Evaluating collaborative filtering recommender systems *ACM Transactions on Information Systems* **22(1)** 5-53
[29] Zhou Y, Wilkinson D, Schreiber R and Pan R 2008 Large-scale parallel collaborative filtering for the Netflix prize *Proc. of Int. Conf. on Algorithmic Applications in Management* 337-348

Acknowledgements
The paper was prepared within the framework of the Academic Fund Program at the National Research University Higher School of Economics (HSE) in 2019 (grant 19-04-0004) and by the Russian Academic Excellence Project 5-100.