Recommendations Based on Visual Content

Taras Hnot

National University of Life and Environmental Science of Ukraine, Department of Economic Cybernetics
tarashnot@gmail.com

Abstract. There is a large number of algorithms to perform recommendations for customers of online platforms. All depends on the data sources we have. Widely used approaches are based on transactional data and "ratings" matrices. For such kind of products as clothes, furniture, hand clocks it is very important to take into account not only some metadata characteristics, but also their "visual look". People always buy clothes based not only on their size, sleeves lengths, textile type, etc., but based on how it looks in general. In this poster paper, we will show how feature vectors of visual content could be extracted and used to enhance recommendations.

Keywords: Visual Recommendations, Deep Neural Networks, ResNet50, Deep features representation, fine-tuning of NN.

1 Introduction

In today's world, there are multiple ways to perform recommendations starting from using attributes and metadata of the products and ending with rates, received by multiple users. In this paper, we are proposing to include into recommendations also visual information, which could be extracted from photos of the products.

Visual information is stored in pixel values of the images. But exact pixels' representation is not the best way to represent images' features. These values are shifted towards position of the object on the image, lighting, etc. It is better to use some "deeper" representation, which could be extracted using neural networks.

2 Neural Networks for Deep Features Extraction

The best way to extract features of images is to use some neural network, which uses these features to perform classification. Earlier layers of such networks give us an opportunity to represent images in the best possible way for comparison with nearest neighbors.

Fine-tuning [3] is a frequently used approach while training neural networks with images. The main idea is to use already trained model and only slightly tune it to work with new data of the same nature. This approach is very useful while working with limited number of data. For example, to train image classifier from scratch we need tens of thousands of observations per class and days of training to achieve high accuracy. In case of fine-tuning it will be enough to have just few hundreds of images per class and a model could be trained in just a few minutes. This could be achieved by using pretrained deep features and building even linear classifier on top of them.

To create a model, subset of DeepFashion[2] dataset was used (46,985 images) to train 46-classes classifier (shirt, cutoff, jeans, suit, etc.). Subset of it could be seen on Fig.1.

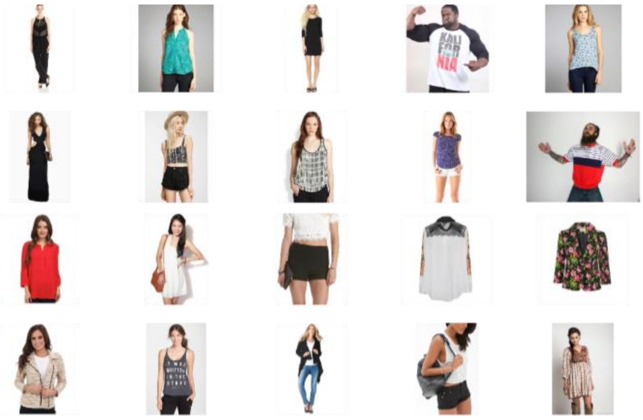

Fig. 1. Sample of DeepFashion dataset

To do fine-tuning, some base model is needed. For that purpose we have used ResNet5[1] trained on ImageNet (1000 classes of 1.28 millions of images).

Process of training is next:

1. Cut off last output layer with 1000 neurons;
2. Add two fully-connected layers (256, 64 neurons) with RELU activation;
3. Add output layer with 46 possible outputs with SOFTMAX activation;
4. Freeze all weight except just added;
5. Train new weights for 10 epochs with ADAM optimizer;
6. Unfreeze all other weights;
7. Fine-tune all weights for 10 epochs with very small learning rate, like (0.001).

Following approach described above, we have achieved 0.76 top 3 accuracy (top 3 means that observation is classified correctly if true value is predicted in top 3 classifier’s outputs).

Then deep image features could be extracted from network using activation of layer before two last layers, which perform classification. In our case – they are 1000 numeric vectors. After the whole dataset of ~290 thousands of images was processed to extract features vectors, comparison was performed using Euclidean distance. Achieved results are on Fig.2.

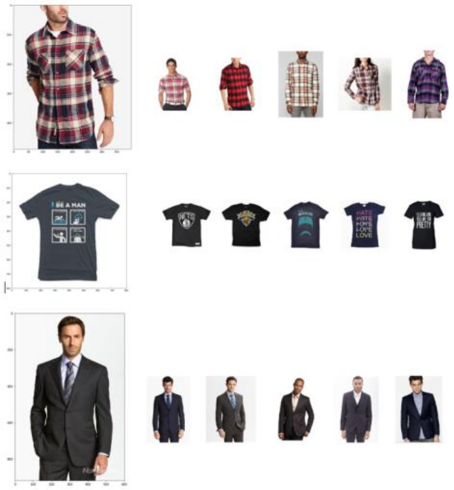

Fig.2. Visual recommendations (first column – input images, next columns – visually similar, sorted by distance, starting from the closest one)

These kind of recommendations are not final. They could be improved by incorporating into feature vector information, related to e.g., color, style, patter.

3 Conclusions

Deep features give an ability to extract information from a visual content that is important for specific task. In our work, we have showed that models, which are used to extract these features, could be trained easily on small data sets using such technique as fine-tuning.

References

1. Kaiming He, Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016).
2. Ziwei Liu, Sijie Yan, Ping Luo, Xiaogang Wang, Xiaoou Tang: DeepFashion: Powering Robust Clothes Recognition and Retrieval with Rich Annotations. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, (2016).
3. Ec Lin, Z., Ji, K., Kang, M., Leng, X., Zou, H.: Deep Convolutional Highway Unit Network for SAR Target Classification with Limited Labeled Training Data. IEEE Geosci. Remote Sens. Lett. 14, 1091–1095 (2017).