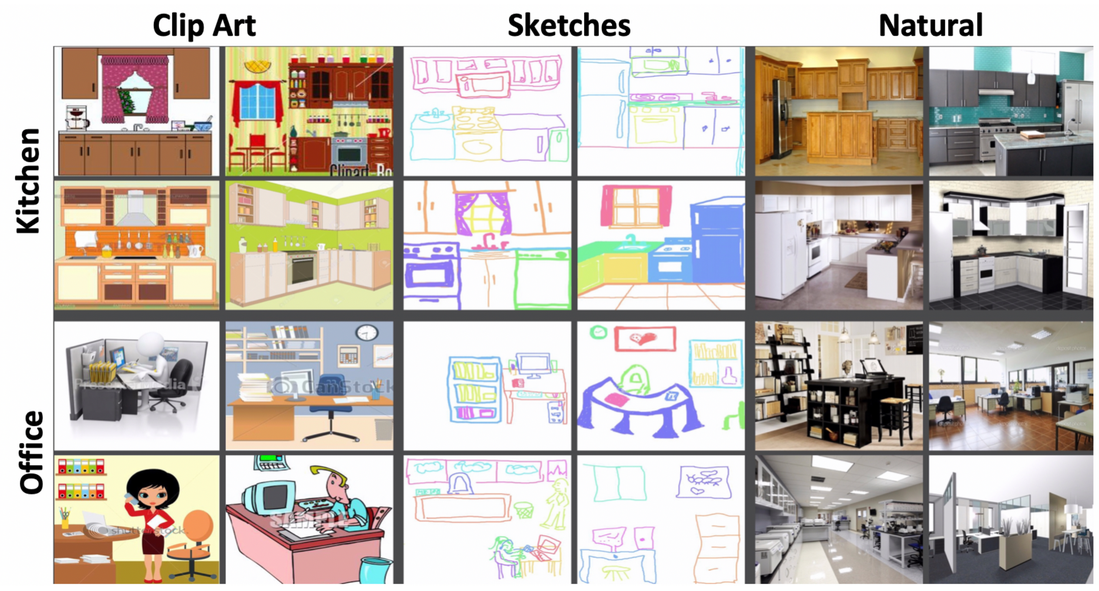
Multi-Source Domain Adaptation: From Natural Images to Clip Art and Sketches

Marius Marten Kästingschäfer

Faculty of Psychology and Neuroscience, Maastricht University, Maastricht, The Netherlands

Can you recognize scenes across stlyes?

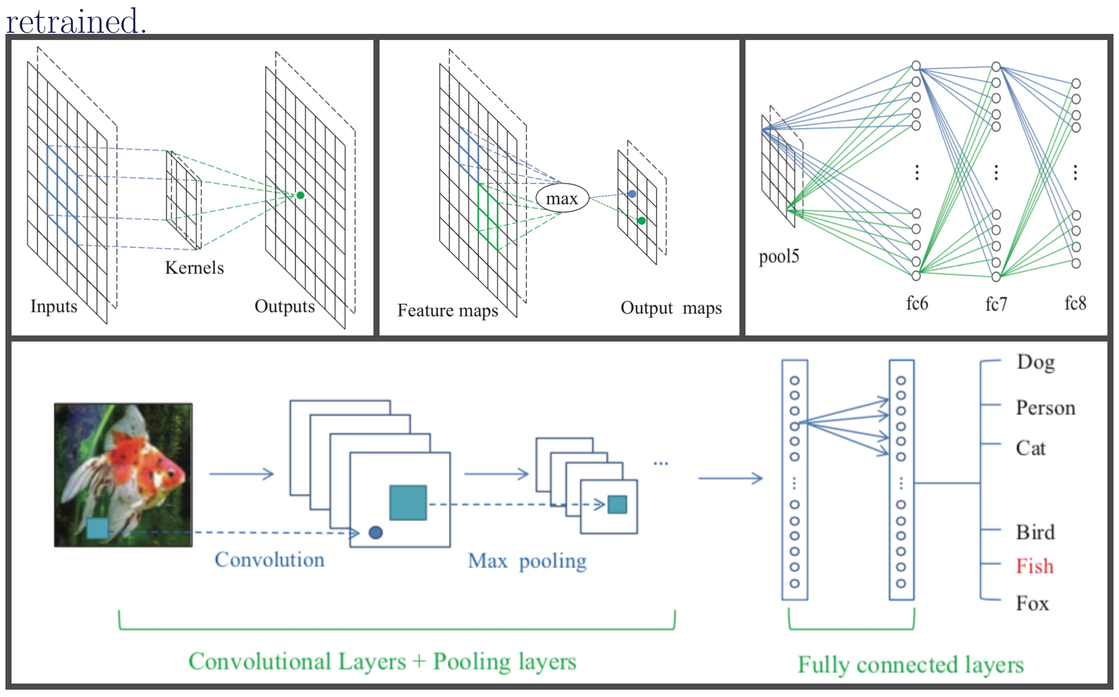


Objective

Humans are able to recognize scenes independently of the **modality** they perceive it in. In this paper, it is tested if scene specific features extracted from natural images are reusable for classifying clip art and sketches. Further, the ability of networks to hold **multiple representations** simultaneously is examined.

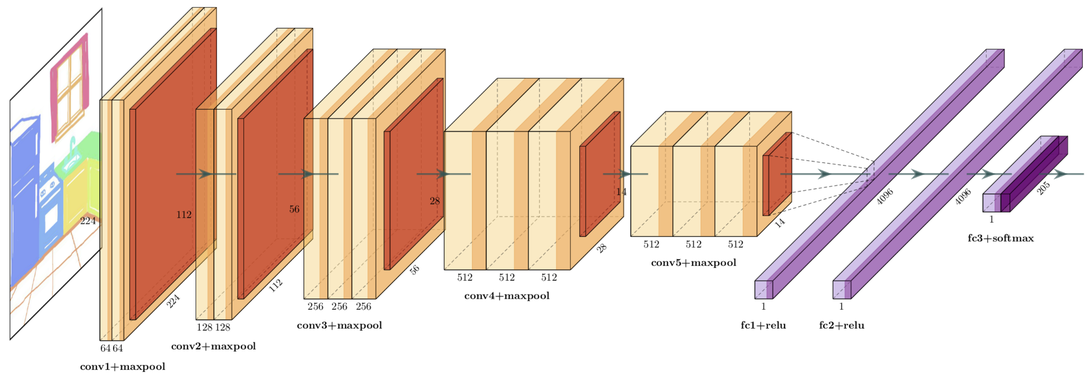
Methods

To study this problem a pretrained **convolutional neural network** is used. A VGG16 architecture [1] was used, consisting of 14.714.688 million parameters partioned over 13 convolutional layers. The 5 maxpooling layers and the ReLU which is applied after every convolutional layer, do not add further parameters. A **randomly initialized classifier** is added and retrained.



The picture [3] in the upper left corner shows the **convolutional operation**. The picture in the upper middle shows the **max pooling** operation and the picture on the upper right shows a schematic representation of the **fully connected layers**. The bottom picture summarizes the mechanisms and is a simplification of the figure overall network. The training consists of adjusting weights to optimally map the 4D image tensor to the 1D labels.

The image classification architecture was trained with **back-propagation** and **stochastic gradient descent**. The learning rate was set to $lr = 1e^{(-4)}$ (0.0001). It took an overall training time of 181 hours (7.5 days) to obtain the result.



Showing the overall **VGG16 architecture** with the randomly initialized classifier in purple. The **datasets** for natural images, clip art and drawings each consisted of around 16.400 pictures.

Results

The representations stored within the pretrained ConvNet are reusable for classifying clipart and sketches. All models performed best on the domain they were trained on. The representations of **clip art** achieve the highest results, followed by **sketches**. **Natural image** performance remains notably below prior results.

Name of the Results	Accuracy
Natural	16.150%
Natural-Sketch	1.797%
Natural-Clip Art	4.973%
Sketches	18.380%
Sketches-Natural	1.386%
Sketches-Clip Art	6.623%
Clip Art	32.000%
Clip Art-Sketch	4.083%
Clip Art-Natural	2.566%

Table 1: Different training-test combinations and their accuracy values

Conclusion

The experiment suggests that the overall transferability of learned features is not only limited by the distance but also by the diversity of the input distribution. Moreover, a multimodal representation is shown to be feasible but only poorly. Further research is needed to explain the obtained outcomes and to explain why they deviate from previous expectations.

References

[1] L. Castrejón, Y. Aytar, C. Vondrick, H. Pirsiavash, and A. Torralba. Learning aligned cross-modal representations from weakly aligned data. *CoRR*, abs/1607.07295, 2016. doi: 10.1109/CVPR.2016.321.

[2] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, 2014.

[3] G. Yanming. *Deep Learning for Visual Understanding*. PhD thesis, ASCI graduate school TU Delf, 2017.

Acknowledgements

I thank Alexander Kroner for helpful discussions and guidance, as well as the Maastricht Research Based Learning (MaRBLe) program. I would also like to thank Salomé-Marie Porten for providing me with continuous encouragement.

Contact Information

- marius.kaestingschaefer@online.de
- http://mariusmarten.github.io