Optimal Input Scale Transformation Search for Deep Classification Neural Networks

Maksim Penkin[1], Alexander Khvostikov[1] and Andrey Krylov[1]

[1]*Faculty of Computational Mathematics and Cybernetics, Lomonosov Moscow State University, Leninskie Gory, 1, building 52, Moscow, 119991, Russia*

Abstract

The paper deals with problem of optimal input scale search for deep classification neural networks. It is shown that state-of-the-art deep neural networks are not stable to input image scale, leading to quality degradation. The paper demonstrates relevance of the topic on classical image classification DL-pipeline. Unlike previous researchers, who aim to build entire complex invariant neural nets, we claim that computing optimal input transformations (e.g. scale) is a more perspective way for successful neural networks real-life applications. Thus, a new scale search algorithm for DL image classification is proposed in the paper, based on empirical hierarchical analysis of activation values.

Keywords

Image scale estimation, Deep learning, Image classification, Medical imaging.

1. Introduction

Many classical computer vision tasks [1, 2] have achieved a great breakthrough, primarily due to the large amount of training data and by the reason of deep learning (DL) application. In recent years, computer vision has been significantly advanced by the adoption of convolutional neural networks (CNNs), so we are currently witnessing many CNN-based models revealing state-of-the-art results in many vision tasks, including image recognition [3], semantic segmentation [4], image captioning [5], etc.

Evolving this progress, there are studies trying to understand what CNNs learn internally [6] and what contribute to its success. By design, layers within neural network have progressively larger receptive field, allowing them to learn more complex features. The key point is the shift invariance property, that a pattern in the input can be recognized regardless of its position. Pooling layers contribute resilience to slight deformation as well as minor scale change. However, CNNs deal with scale variance far worse than shift variance [7]. Not dealing with scale invariance well poses a direct conflict to the design philosophy of CNN, in that higher layers may see and thus capture features of certain plain patterns simply because they are larger at the input, not because they are more proper.

Many recent works have focused on introducing transformation invariance in deep learning architectures explicitly.

For unsupervised feature learning, a transform invariant restricted Boltzmann machine is presented [8] that compactly represented data by its weights and their transformations, which achieved invariance of the feature representation via probabilistic max pooling.

Multi-scale learning techniques are proposed in [9, 10], where each CNN is trained over multiple scales independently without weight sharing.

Another work [11] proposed raw input image to be transformed through a Laplacian pyramid. Each scale was fed into 3-stage convolutional network, which produced a set of feature maps with different scales disjointly. Then, the outputs of all scales were aligned, by upsampling, and concatenated. However, concatenating multi-scale outputs to extract scale independent features involves extra

GraphiCon 2022: 32nd International Conference on Computer Graphics and Vision, September 19-22, 2022, Ryazan State Radio Engineering University named after V.F. Utkin, Ryazan, Russia
EMAIL: penkin97@gmail.com (M. Penkin); khvostikov@cs.msu.ru (A. Khvostikov); kryl@cs.msu.ru (A. Krylov)
ORCID: 0000−0002−8027−9333 (M. Penkin); 0000−0002−4217−7141 (A. Khvostikov); 0000−0001−9910−4501 (A. Krylov)


convolutional layers and extra parameters, thereby increasing the complexity of the models and the computational cost, increasing the risk of overfitting.

Unlike stated above researchers [8, 9, 10, 11], who aim to train scale invariant deep learning algorithm, once faced poor scale invariance of common CNNs, we desire to view the problem from the other side, searching an optimal input scale transformation to an already pre-trained deep learning algorithm. Optimality of transformation is considered in terms of maximizing prediction accuracy.

As for the field of real applications, the problem of input scale selection is particularly very relevant for histological images.

The advent of digital high-resolution scanners has made available digitized histological tissue samples that are suitable for computer-aided diagnosis (CAD). CAD can relieve the pathologists' work by discriminating obviously benign and malignant tissue, thus reducing the amount of tissue area to be analyzed by a pathologist.

The choice of scale at which perform the image analysis, however, is typically arbitrary. Pathologists usually identify suspicious regions at lower resolutions and only use the information at the higher scales to confirm their suspicions (see Figure 1). Digitized breast cancer histopathology image in Figure 1 contains information at multiple scales. While low level attributes such as texture and intensity are available at the lower image scales to distinguish benign from cancerous regions, higher level shape and architectural attributes of tissue become apparent only at the higher scales.

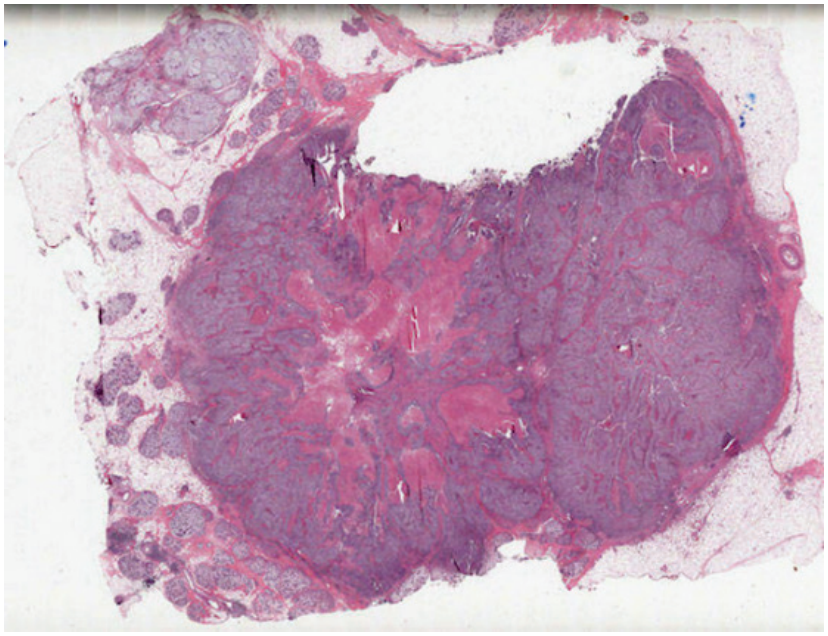

Figure 1: Example of a histological whole-slide image

Deep learning techniques have proved to be efficient computer assisted systems for digital medical image analysis due to their feature representation capabilities. Recently proposed CNNs are successfully applied in the field of medical imaging [12, 13], being a very helpful tool for doctors. For example, the principle is used for instance-based gland segmentation [14, 15, 16].

Nevertheless, it is still not clear what input transformations are optimal for a selected deep neural model to attain the most qualitive and representative output result.

Our method tries to shed the light onto this question, providing the empirical unsupervised transformation (for now, scale) selection rule for a chosen CNN algorithm.

The paper is organized in a following way: Section 2 describes relevance of the proposed algorithm for classification task; Section 3 directly reveals the proposed method; Section 4 reports the conducted classification experiments both ImageNet and medical imaging; Section 5 concludes the research.

2. Relevance

We start our research with topic relevance assessing, trying to answer the question, how stable classification DL-algorithms (CNNs) are to input scale transformations.

Image classification [2], as classical research topic, is one of the core issues of computer vision and the basis of various fields of visual recognition. The success in classification networks is closely connected with other image processing and computer vision fields.

So, in this section we examine the scale invariance property of image classification CNNs, revealing the relevance of the proposed optimal input scale search algorithm for classification pipeline.

The scale sensitivity is evaluated for several state-of-the-art pre-trained architectures: AlexNet [17], VGG19 [18], ResNet18 and ResNet50 [19]. These models had been pre-trained on a large dataset ImageNet [20] and are available now in majority of contemporary DL-frameworks: Keras, TensorFlow, PyTorch. We used PyTorch in conjunction with Python 3 in all our experiments.

Scale invariance was checked by passing ImageNet images through SoTA pre-trained CNNs in various input scales. Figure 2 reveals the instability of different architectures to input scale changes in a range $[-3,\ 4]$, where -3 corresponds to $\times\frac{1}{3}$ scale; {-1, 0, 1} correspond to original ImageNet resolution 224×224; 3 matches $\times$ 3 scale. It could be noted that the considered optimal scale value depends on the architecture: AlexNet is mostly confident in its answer on the training original resolution 224×224, VGG19 trusts $\times\frac{1}{2}$ scale, ResNet50 has no explicit local maximum on the selected scale range.

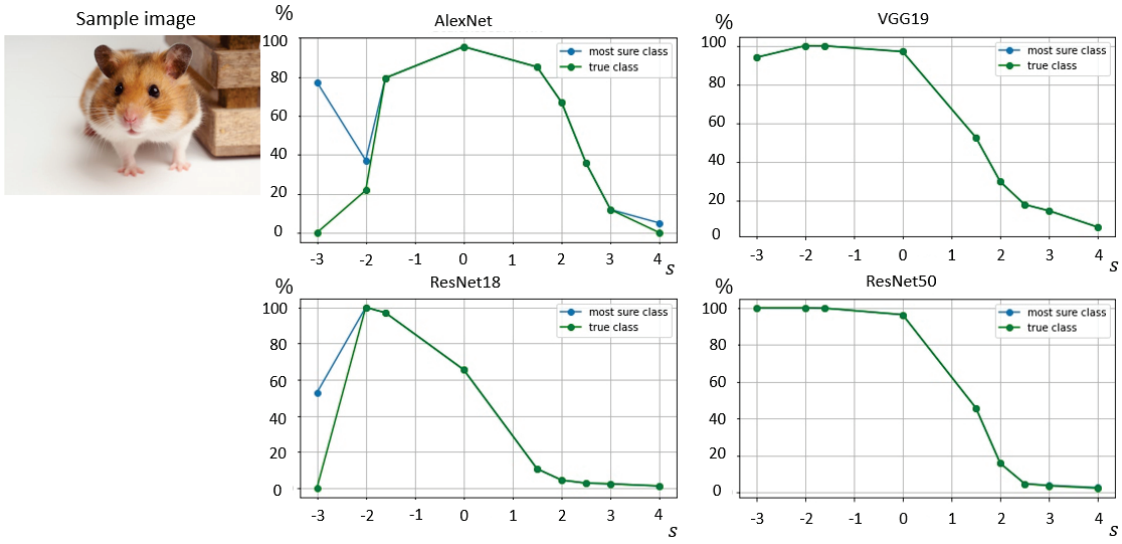

Figure 2: SoTA classification networks prediction stability upon various input scales s of a sample image. Blue line: the most sure class; green line: ground truth class. OX: scale range $s \in [s_{min}, s_{max}]$ (-2 corresponds to $\times\frac{1}{2}$ scale; 2 corresponds to $\times$ 2 scale; $\{-1, 0, 1\}$ match original resolution); OY: classification accuracy (%)

Additionally, it is worth to notice that some nets are more persistent to downscale than upscale: ResNet50 exhibits severe confidence degradation on upscale and almost no confidence distortion on downscale. Furthermore, a special explanation tool for CNNs: Grad-CAM [21], – was utilized to prove classifiers' instability upon varying input scale, and thus proving relevance of the proposed topic. Grad-CAM is an explanation tool that uses the gradient information of the target object and how it flows through a network to create coarse localization heatmaps for visual analysis. The heatmap produced by Grad-CAM (see Figure 3) tells clearly for an image, which parts are under focus and considered by a CNN to come to a decision. Blue parts of the heatmap indicate no participation and red parts indicate high participation.

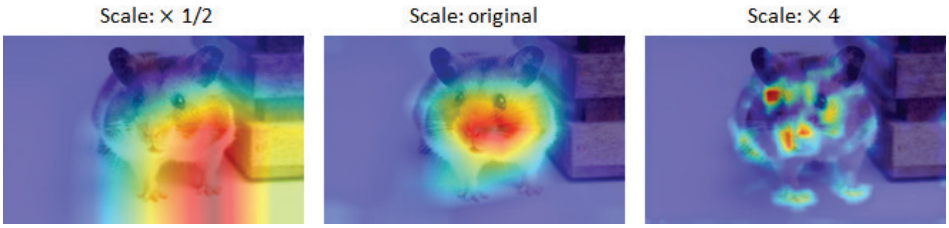

Figure 3: VGG19 Grad-CAM visualization on different input scales: explicit shift of the region of participation is highlighted by the algorithm

3. Proposed method

We propose a new unsupervised algorithm for searching an optimal input data transformation for a chosen CNN. In this paper we fix *input transformation* to *scale* and *CNN* to *VGG19* (see Figure 4) for simplicity.

The proposed algorithm evolves an intuition underlying the correlation filter principle [22] widely used in convolutional neural nets nowadays: pattern is being recognized by correlating the filter over a sliding window; corresponding high output absolute magnitudes indicate a success in pattern recognition. Recently, in 2010's [17] the correlation filter principle was unfolded into the deep non-linear ensemble, named convolutional neural network (CNN).

Consequently, our hypothesis is that the appropriate input scale selection induces activations magnitude hike across a neural network.

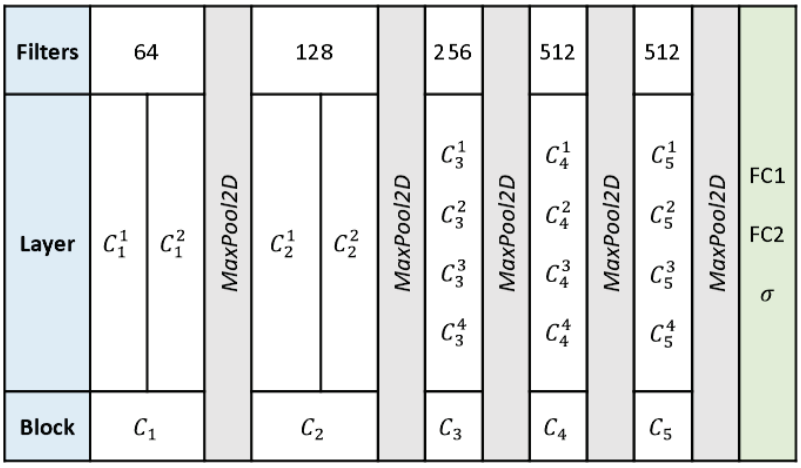

Figure 4: VGG19 architecture

The proposed optimal transformation search algorithm has several constraints:

- feature maps f_k should meet extremum in interior of a scale search range (and it corresponds to the real optimal scale);
- analyzing layer set $\mathcal{C}$ should be chosen a priori.

First constraint is typical for many mathematical algorithms, for example, classical root-finding methods have the same restriction: Newton's method, Secant method.

Second constraint is caused by the recency (novelty) of the proposed approach. Currently, there is no layers' selection rule. Layers are selected for each model independently, based on general deep learning assumptions and intuitions.

Algorithm's description.

1. Select scale search range: $s \in [s_{min}, s_{max}]$.

2. Extract VGG19 layers: $\mathcal{C} = \{C_1^2, C_2^2, C_3^4, C_4^4, C_5^4\}$, where C_i^j corresponds to j-th activated convolutional tensor in i-th VGG19 block.
3. For each $C_i^j \in \mathcal{C}$ do:
    a. Calculate its values as a function $C_i^j(s)$ on the selected scale range: $s \in [s_{min}, s_{max}]$.
    b. Discard monotonous & constant feature maps upon the scale grid by $\mathcal{F}$ filter: $\mathcal{F}: \{f_k(C_i^j)(s), k \in [1..n_i^j]\} \to \{f_k(C_i^j)(s), k \in [1..\tilde{n}_i^j]\}$, where $f_k(C_i^j)$ corresponds to k-th feature map of C_i^j convolutional tensor.
    c. Merge filtered feature maps by $\mathcal{M}$ operator, which is set here as mathematical expected value function ($\mathbb{E}$): $\mathcal{M}: \{f_k(C_i^j)(s), k \in [1..\tilde{n}_i^j]\} \to \hat{C}_i^j(s)$.
    d. Compute layer's scale prediction as a scale, corresponding to a local extremum: $\hat{s}_i^j = \arg \max_{s \in [s_{min}, s_{max}]} \hat{C}_i^j(s)$.
4. Aggregate layers' predictions $\{\hat{s}_i^j\}_{ij}$ via linear regression (minimizing l_2 error) to a single optimal scale value $\hat{s}$.

4. Experiments

The described above scale selection algorithm and the corresponding hypothesis, lying underneath, have been verified for VGG19 network and two classification pipelines: ImageNet and medical histology.

Demo-version of the proposed input scale selection approach is shared on open-source hosting for software development Github[2]. Demo-version is coded using Python 3 and PyTorch deep learning framework. Algorithm is iterative and non-trainable; thus, it can be evaluated both on CPU & GPU, the only restriction is memory: the amount of RAM (system or GPU) should be large enough to store and inference a chosen CNN model.

4.1. ImageNet classification

Feasibility experiments of the proposed method were conducted, firstly, on ImageNet dataset for pre-trained VGG19 classification convolutional neural network.

Feature maps are analyzed for each selected activated convolutional tensor C_i^j according to the algorithm's formulation (see section 3).

Let's consider C_2^2, for example (see Figure 5 (a)). Feature maps of the convolutional tensor $C_2^2 \in \mathcal{C}$ are filtered by $\mathcal{F}$ operator, discarding monotonous and constant ones. For visualization purposes 2D feature maps are projected onto a field of real numbers ($\mathbb{R}$) with Average Pooling DL-operator, yielding *feature value* (see OY of Figure 5 (a)).

Then, sample mean (or "empirical mean") is calculated over feature maps, thus, merging ($\mathcal{M}$) them together in a single curve (see Figure 5 (b)).

C_2^2's prediction $\hat{s}_2^2$ is extracted as a scale, corresponding to a local extremum of a merged curve (see Figure 5 (b)).

The final scale prediction $\hat{s}$ is calculated as linear regression of layers' predictions $\{\hat{s}_i^j\}_{ij}$ (see Figure 6 (a)). The derived $\hat{s}$ corresponds to near-optimal VGG19 accuracy (see Figure 6 (b)).

Scale search method was tested on ImageNet testing subset (120 randomly chosen testing images). Several results can be seen in Figure 10, attached to Appendix section. Low variance of layer's (C_i^j's) predictions can be noted, however, for some particular samples, algorithm's confidence is low, as layers' predictions $\{\hat{s}_i^j\}_{ij}$ are noisy.

[2] https://github.com/MaksimPenkin/ScaleEstimation

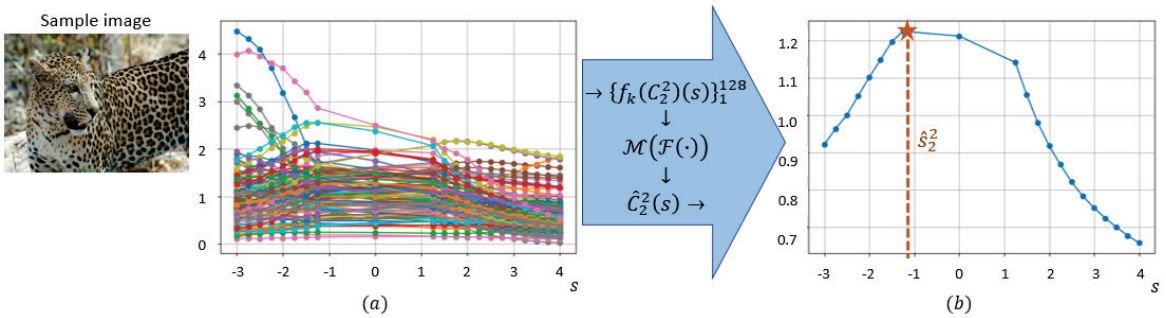

Figure 5: Proposed algorithm visualization for VGG19 C_2^2 layer. (a) – $\{f_k(C_2^2)(s), k \in [1..128]\}$ feature maps dependencies upon various input scales s; (b) – merged feature curve $\hat{C}_i^j(s)$. OX: scale range $s \in [s_{min}, s_{max}]$ (-2 corresponds to $\times\frac{1}{2}$ scale; 2 corresponds to $\times$ 2 scale; $\{-1, 0, 1\}$ match original resolution); OY: feature value

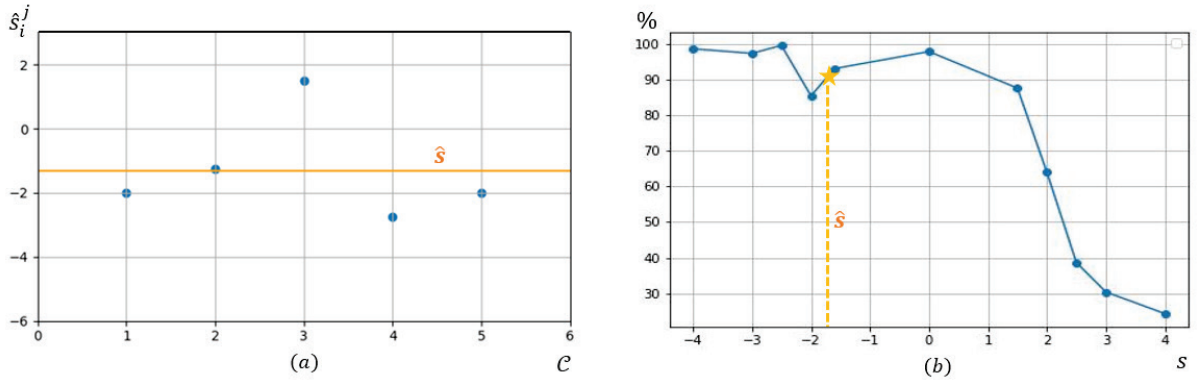

Figure 6: (a) – Linear aggregation of algorithm's layers' predictions $\{\hat{s}_i^j\}_{ij}$ for a sample image; (b) – VGG19 accuracy dependency upon input scales s

Notable is the fact that percentage of selected feature maps by $\mathcal{F}$ operator is gradually raising to almost 100% with increasing VGG19's depth (see Figure 9, attached to Appendix section).

Another remarkable point is that on 63% of ImageNet testing images VGG19 is *more or equal* ($\geq$) confident in corresponding ground truth class with proposed scale rather than with original one, and on 14% of testing images VGG19 is strictly *more* ($>$) sure in corresponding true class with our proposed scale (see Figure 11, attached to Appendix section).

4.2. Histology feature extraction

Algorithm's applicability was also checked for histological image classification. As mentioned above, input scale selection issue is specifically relevant for histological image processing.

Experiments were carried out for pre-trained VGG19. Basically, there are several ways of using pre-trained classification neural architectures: as a baseline for further fine-tuning or as a feature extraction tool. Here, VGG19 neural network was utilized as a feature extraction tool, the same used authors in [23], where fine-tuning approach was reported to demonstrate not good enough medical classification performance. So, the question is which scale of a histological whole-slide image (see Figure 7) induces the most extensive response of VGG19 neural network – non-linear ensemble of correlation filters.

The proposed algorithm produces the following dependencies $\{\hat{C}_i^j(s)\}_{ij}$ upon input scales (see Figure 8). It can be noted that first VGG19 blocks (C_1, C_2) bring out strongly convex curves, revealing single optimal scales $\hat{s}_1^2, \hat{s}_2^2$, respectively, as solutions. However, deep VGG19 blocks C_3, C_4, C_5 yield almost monotonic increasing functions, not allowing the proposed algorithm to resolve the solution.

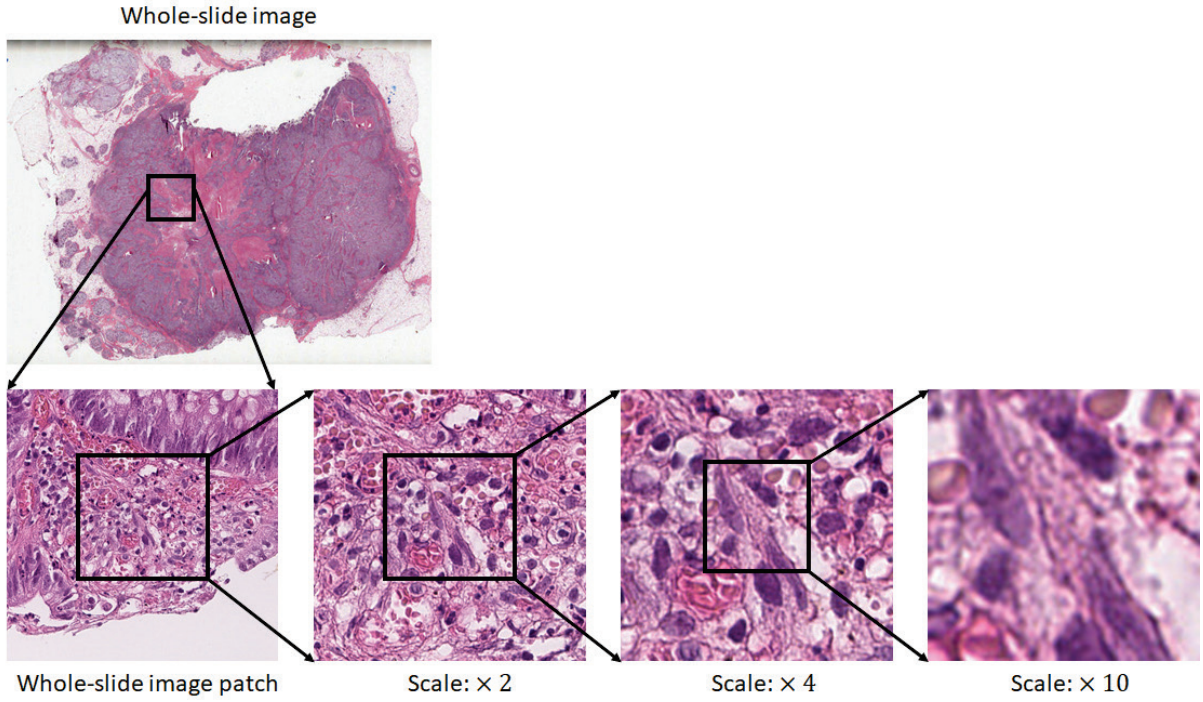

Figure 7: Various scales of a histological whole-slide image

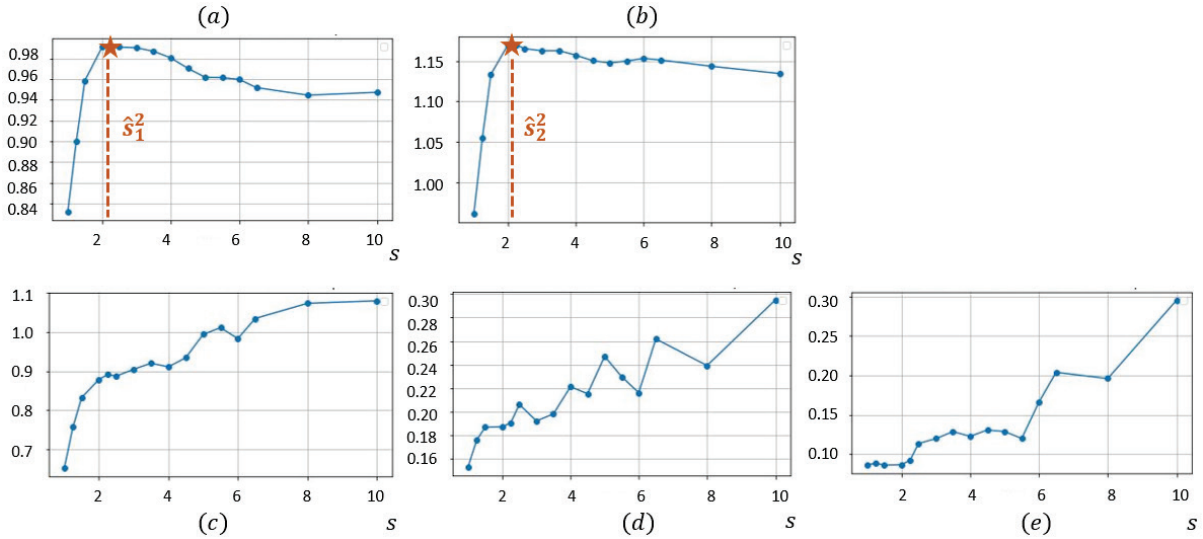

Figure 8: $(a) - \hat{C}_1^2(s)$; $(b) - \hat{C}_2^2(s)$; $(c) - \hat{C}_3^4(s)$; $(d) - \hat{C}_4^4(s)$; $(e) - \hat{C}_5^4(s)$. OX: scale range $s \in [s_{min}, s_{max}]$ (2 matches × 2 scale; 10 matches to × 10 scale); OY: feature value

5. Conclusion

In this paper we presented the optimal input transformation (scale) selection algorithm. The proposed algorithm is new, empirical, unsupervised, iterative and non-trainable approach.

Relevance of the developed method is obvious for histological image analysis, furthermore we showed importance of such research for the classical classification pipeline, as quality of common image analysis acutely depends on correct input transformations.

The revealed algorithm has weaknesses, which are planned to be eliminated in the course of future research. Main issues to be enhanced are feature maps merging procedure and layers' predictions aggregating strategy. Merging feature maps by sample mean statistic may not be good due to its instability to outliers. Another concern is usage of standard linear regression, which gives interpretable solution, yet being not a robust tool.

The obtained algorithm has shown its potential applicability and will be further developed, especially for histological image analysis.

6. Acknowledgements

The work was supported by RSCF grant 22-41-02002.

7. References

[1] X. Liu, L. Song, S. Liu, Y. Zhang, A review of deep-learning-based medical image segmentation methods., Sustainability 13(3) (2021) 1224. doi:10.3390/su13031224.

[2] A. K. Sharma, A. Nandal, A. Dhaka, R. Dixit, Medical image classification techniques and analysis using deep learning networks: a review., Health Informatics: A Computational Perspective in Healthcare (2021) 233–258. doi:10.1007/978-981-15-9735-0_13.

[3] B.B Traore, B. Kamsu-Foguem, F. Tangara, Deep convolution neural network for image recognition., Ecological Informatics 48 (2018) 257–268. doi:10.1016/j.ecoinf.2018.10.002.

[4] Y Guo, Y. Liu, T. Georgiou, M. S. Lew, A review of semantic segmentation using deep neural networks., International journal of multimedia information retrieval 7(2) (2018) 87–93. doi:10.1007/s13735-017-0141-z.

[5] J. A. Alzubi, R. Jain, P. Nagrath, S. Satapathy, S. Taneja, P. Gupta, Deep image captioning using an ensemble of CNN and LSTM based deep neural networks, Journal of Intelligent & Fuzzy Systems 40(4) (2021) 5761–5769. doi:10.1109/ICME.2017.8019408.

[6] S. Mostafa, D. Mondal, M. Beck, C. Bidinosti, C. Henry, I. Stavness, I., Visualizing feature maps for model selection in convolutional neural networks.,In Proceedings of the IEEE/CVF International Conference on Computer Vision (2021) 1362–1371. doi:10.1109/ICCVW54120.2021.00157.

[7] Y. Gong, L. Wang, R. Guo, S. Lazebnik, Multi-scale orderless pooling of deep convolutional activation features., In European conference on computer vision (2014) 392–407. doi:10.1007/978-3-319-10584-0_26.

[8] K. Sohn, H. Lee, Learning invariant representations with local transformations., arXiv preprint arXiv:1206.6418 (2012).

[9] J. M. Alvarez, Y. LeCun, T. Gevers, A. M. Lopez, Semantic road segmentation via multi-scale ensembles of learned features., In European Conference on Computer Vision (2012) 586–595. doi:10.1007/978-3-642-33868-7_58.

[10] P. Sermanet, Y. LeCun, Traffic sign recognition with multi-scale convolutional networks., In The 2011 international joint conference on neural networks (2011) 2809–2813. doi:10.1109/IJCNN.2011.6033589.

[11] C. Farabet, C. Couprie, L. Najman, Y. LeCun, Learning hierarchical features for scene labeling., IEEE transactions on pattern analysis and machine intelligence 35(8) (2012) 1915–1929. doi:10.1109/TPAMI.2012.231.

[12] A. Sinha, J. Dolz, Multi-scale self-guided attention for medical image segmentation., IEEE journal of biomedical and health informatics 25(1) (2020) 121–130. doi:10.1109/JBHI.2020.2986926.

[13] H. Huang, L. Lin, R. Tong, H. Hu, Q. Zhang, Y. Iwamoto, X. Han, Y. W. Chen, J. Wu, Unet 3+: A full-scale connected unet for medical image segmentation., In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2020) 1055–1059. doi:10.1109/ICASSP40776.2020.9053405.

[14] M. Salvi, M. Bosco, L. Molinaro, A. Gambella, M. Papotti, U. R. Acharya, F. Molinari, A hybrid deep learning approach for gland segmentation in prostate histopathological images., Artificial Intelligence in Medicine 115 (2021) 102076. doi:10.1016/j.artmed.2021.102076.

[15] A. Khvostikov, A. S Krylov, I. Mikhailov, P. Malkov, CNN Assisted Hybrid Algorithm for Medical Images Segmentation., In Proceedings of the 2020 5th International Conference on Biomedical Signal and Image Processing (2020) 14–19. doi:10.1145/3417519.3417557.
[16] N. Oleynikova, A. Khvostikov, A. Krylov, I. Mikhailov, O. Kharlova, N. Danilova, P. Malkov, N. Ageykina, E. Fedorov, Automatic glands segmentation in histological images obtained by endoscopic biopsy from various parts of the colon., Endoscopy 51(04) (2019) OP9. doi:10.1055/s-0039-1681188.
[17] A. Krizhevsky, I, Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks., Advances in neural information processing systems 25 (2012). doi:10.1145/3065386.
[18] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition., arXiv preprint arXiv:1409.1556 (2014). doi:10.1109/ACPR.2015.7486599.
[19] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition., In Proceedings of the IEEE conference on computer vision and pattern recognition (2016) 770–778. doi:10.1109/CVPR.2016.90.
[20] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database., In 2009 IEEE conference on computer vision and pattern recognition (2009) 248–255. doi:10.1109/CVPR.2009.5206848.
[21] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization., In Proceedings of the IEEE international conference on computer vision (2017) 618–626. doi:10.1109/ICCV.2017.74.
[22] D. S. Bolme, J. R. Beveridge, B. A. Draper, Y. M. Lui, Visual object tracking using adaptive correlation filters., In 2010 IEEE computer society conference on computer vision and pattern recognition (2010) 2544–2550. doi:10.1109/CVPR.2010.5539960.
[23] A. Rakhlin, A. Shvets, V. Iglovikov, A. A. Kalinin, Deep convolutional neural networks for breast cancer histology image analysis., In international conference image analysis and recognition (2018) 737–744. doi:10.1101/259911.

8. Appendix

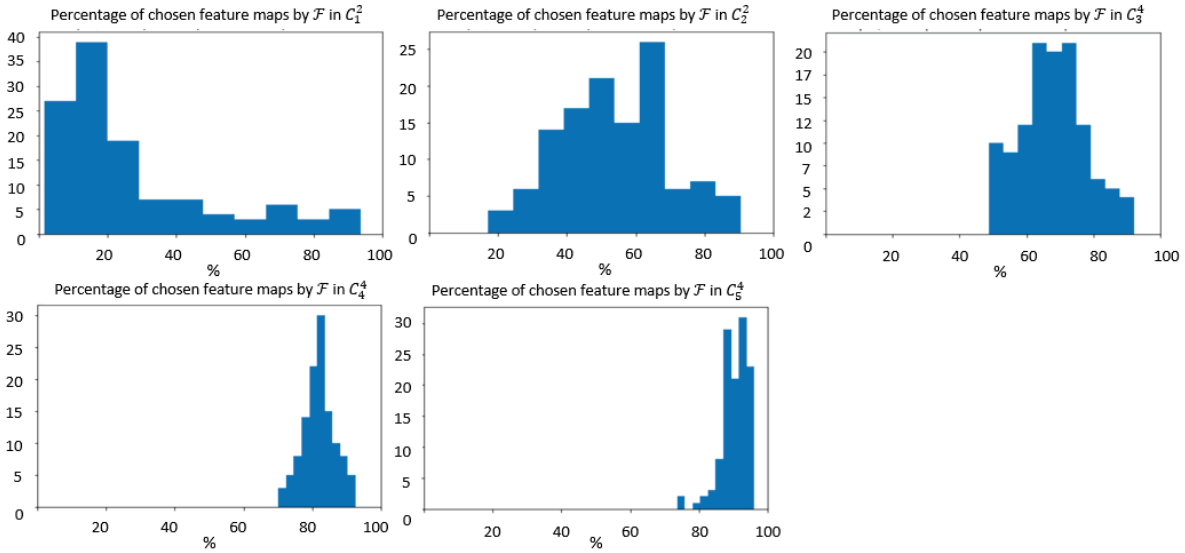

Figure 9: Proposed algorithm exploration on ImageNet testing subset (120 images). Percentage of selected feature maps by $\mathcal{F}$ operator on different VGG19 depth. Histograms are computed on ImageNet subset

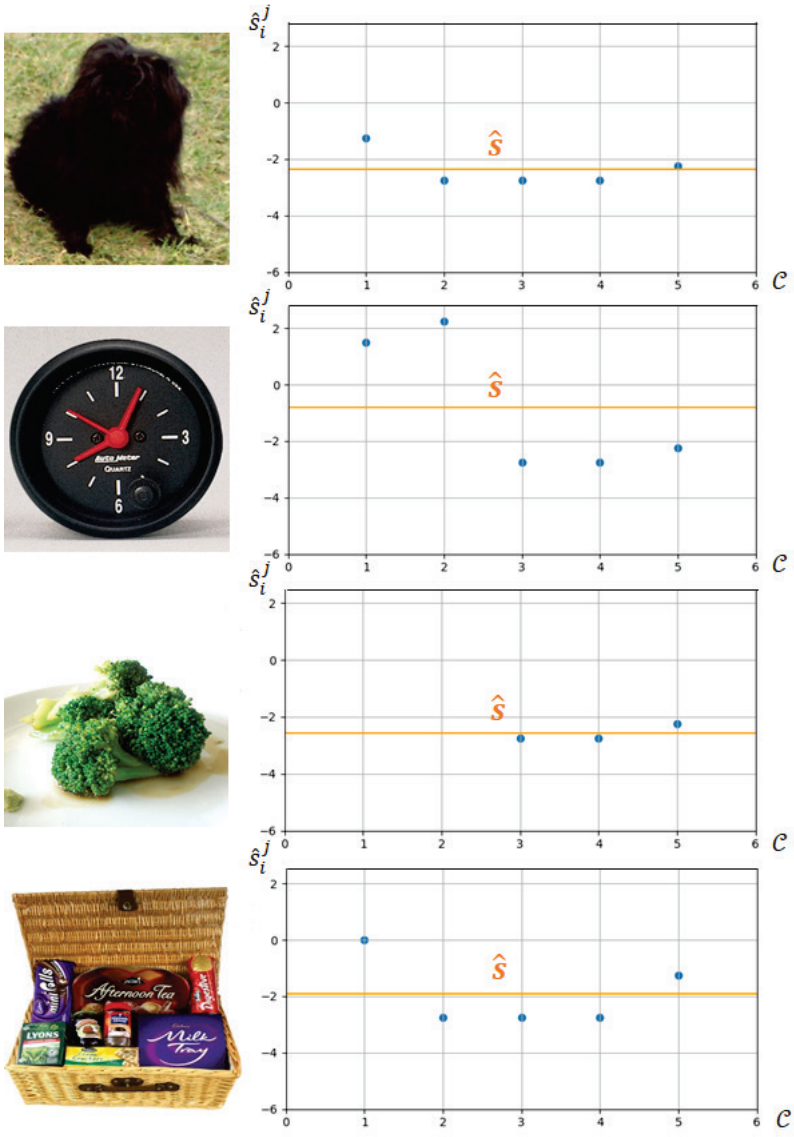

Figure 10: Proposed algorithm evaluation on 4 images from ImageNet testing subset (120 images). Blue dots: $\{\hat{s}_i^j\}_{ij}$ predictions for each $C_i^j \in \mathcal{C}$; orange line: derived algorithm's optimal scale solution $\hat{s}$. OX: i-th VGG19 block, used for computing C_i^j layer prediction $\hat{s}_i^j$; OY: layer's prediction $\hat{s}_i^j$

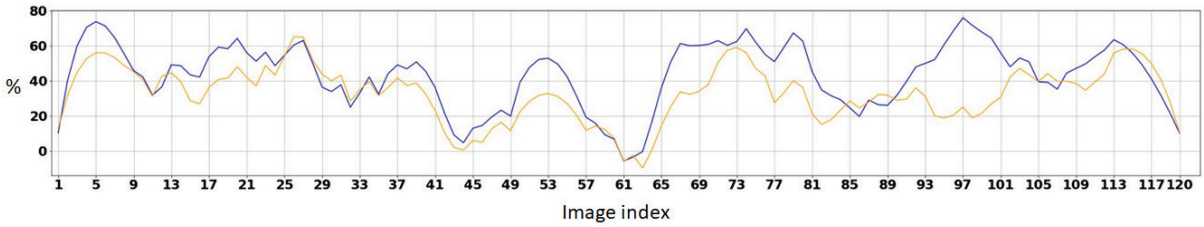

Figure 11: Proposed algorithm evaluation on 120 images from ImageNet testing subset. Blue line: VGG19 predictions on original resolution 224×224; orange line: VGG19 prediction on proposed scale. OX: image index; OY: VGG19 confidence (%) in the ground truth class for each image