MEDINFO 2021: One World, One Health – Global Partnership for Digital Innovation
P. Otero et al. (Eds.)

doi:10.3233/SHTI220263

Comparison of Supervised and Self-Supervised Deep Representations Trained on Histological Images

Dawid Rymarczyk[a,b], Adriana Borowa[a,b], Anna Bracha[b], Maurycy Chronowski[b], Wojciech Ozimek[b], Bartosz Zieliński[a,b]

[a] *Faculty of Mathematics and Computer Science, Jagiellonian University, Lojasiewicza 6, 30-348 Kraków, Poland*
[b] *Ardigen SA, Podole 76, 30-394 Kraków, Poland*

Abstract

Self-supervised methods gain more and more attention, especially in the medical domain, where the number of labeled data is limited. They provide results on par or superior to their fully supervised competitors, yet the difference between information coded by both methods is unclear. This work introduces a novel comparison framework for explaining differences between supervised and self-supervised models using visual characteristics important to the human perceptual system. We apply this framework to models trained for Gleason score and conclude that self-supervised methods are more biased toward contrast and texture transformation than their supervised counterparts. At the same time, supervised methods code more information about the shape.

Keywords:

Deep Learning; Image Processing, Computer-Assisted; Prostatic Neoplasms

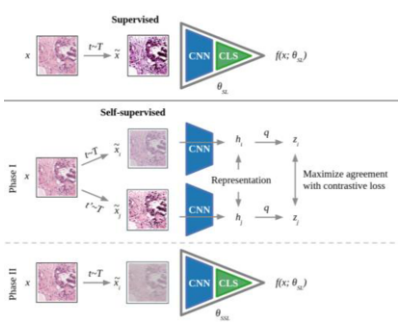

Figure 1 - Supervised learning (SL) and Self-supervised learning (SSL) frameworks.

Introduction

With the growth of deep learning in computer vision applications of neural networks to medical imaging are common. Most of them take the model pretrained on natural images and fine-tune it in a supervised manner (SL) [7]. However, such a pipeline focuses on visual features of natural images, not optimal for the medical domain. Hence, self-supervised learning (SSL) method was introduced [4], which uses unlabeled medical data to learn an adequate representation. Such representations are used with a small number of labeled data and achieve state-of-the-art results [3].

In this work, we introduce the comparison framework for identifying the differences between SL and SSL. The tool is based on SVCCA [6] and image features relevant from the human perspective, identified by neuroscience research [5].

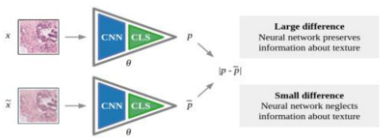

Figure 2 – Comparison Framework

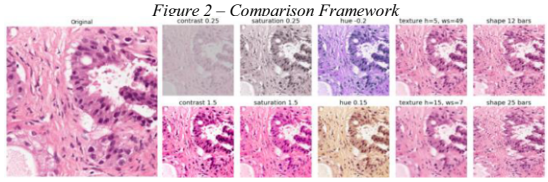

Figure 3 - Sample transformations from our comparison framework.

The main contributions of this work are as follows: (1) Introducing the comparison framework that identifies the differences between representations trained with supervised and self-supervised models. (2) Presenting the difference between SL and SSL models using image features relevant from the human vision perspective. (3) Analyzing the influence of model size on features trained by supervised and self-supervised models.

Methods

Supervised learning

The supervised learning (SL) approach is defined by $y = f(x; \theta_{SL})$, where $x \in X$ is an input from a training set, $y \in Y$ is a corresponding output, and f is a model with parameters θ_{SL} (see Fig. 1). Ground truth values of y are labels in the classification problem or real values in the regression.

Self-supervised learning

Self-supervised learning (SSL) in vision tasks mostly utilizes contrastive methods. In our work, we explore the SimCLR framework [3] which uses multiple data augmentation operations, positive and negative pairs of images, and NT-Xent loss. Firstly, the CNN is trained to provide a task-independent representation, used in the second phase as the input for the classifier (see Fig. 1). Moreover, the first phase uses additional layers q, called projection head, removed after training.

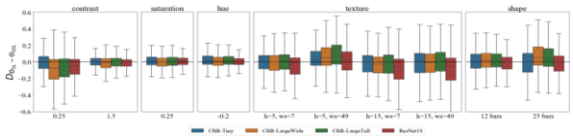

Figure 4 - Distribution of differences between the importance of visual characteristics in supervised and self-supervised methods (all significantly different, with p-values<0.05).

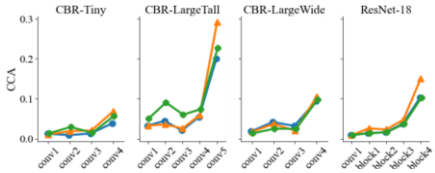

Figure 5 - CCA similarity between model: SL – Ssupervised, SSL – Self-supervised, RI – randomly initialized

Comparison Framework

Human-friendly visual characteristics

According to [5], the most important visual characteristics are customizable, e.g., contrast, saturation, hue, texture, and shape. The sample image and its transformations are in Fig. 3: contrast and saturation are modified using a gray-scale image version, hue is transformed in HSV, non-local means denoising [1] modifies texture, and shapes in an image are modified by shifting its parts. Let Θ_{SL} and Θ_{SSL} be the parameters of the models trained with SL and SSL, respectively, $f(x;\Theta)$ is the prediction of the model with weights Θ for input $x \in X$, where X is the validation set. For each model, we compute set D_Θ of absolute differences between predictions for the original x and their transformations $\bar{x}$: $D_\Theta = \{D_\Theta(x) = |f(x;\Theta) - f(\bar{x};\Theta)| : x \in X\}$. This process is presented in Fig. 2. By applying this step to both models, we obtain two sets of measurements, $D_{\Theta_{SL}}(x)$ and $D_{\Theta_{SSL}}(x)$. Finally, we compute the distribution of differences: $D_{\Theta_{SL}-\Theta_{SSL}} = \{D_{\Theta_{SL}}(x) - D_{\Theta_{SSL}}(x) : x \in X\}$, and use the Wilcoxon signed-rank test to examine symmetricity about zero. If such a null hypothesis is rejected, we find differences between compared models.

Canonical correlation analysis

To analyze latent representations throughout models, we apply Singular Vector Canonical Correlation Analysis (SVCCA) [6]. Firstly, SVD finds the most important directions in latent spaces of the two models, and then CCA aligns them and calculates the correlation between two layers. The higher the CCA value, the strong resemblance between the models.

Results and Discussion

Models' performance

Mean accuracies and standard errors, reported on the hold-out for four trained models, are in Table 1. We observe that the SSL training results are more effective for all setups, more visibly for CBR than for ResNet-18, probably due to the optimal design of the latter. Moreover, the SL works similarly for all networks.

Supervised vs. self-supervised comparison

Image features analysis

Fig. 4 presents distribution of differences between $D_{\Theta_{SL}}$ and $D_{\Theta_{SSL}}$ for the set of transformations. Positive value means the visual characteristic is better represented in the SL, otherwise in the SSL. Results suggest that SSL puts attention on the image contrast and there are no differences for saturation and hue. Interestingly, SL codes information about fine ($h = 5$ and $ws = 49$), while SSL focuses more on coarse textures ($h = 15$). On the other hand, big changes in shape ($n_{bars} = 25$) have a higher impact on SL. Therefore, we conclude that in a case of significant modifications, SSL methods are biased toward contrast and texture, while SL methods focus on the shape.

CCA similarity

Results of the CCA analysis presented in Fig. 5 show that models are rather not similar except for the last layers, responsible for representing high-level features like objects or their significant parts.

Conclusions

We introduced a comparison framework to present differences between the same model trained in different regimes, using visual characteristics important to the human perceptual system. We applied this framework to models trained on the PANDA dataset and found that there are significant differences between features trained by SL and SSL. The SL focues more on the shape and fine texture, while SSL is biased toward contrast and coarse texture.

Table 1 - Effectiveness of trained models over four. Bold values are significantly higher based on Wilcoxon signed-rank test

Architecture	# of Params.	Self-supervised		Supervised	
		Accuracy	AUC	Accuracy	AUC
CBR-Tiny	1.1 M	**0.893 ± 0.015**	**0.960 ± 0.011**	0.851 ± 0.004	0.932 ± 0.002
CBR-LargeWide	8.4 M	**0.896 ± 0.004**	**0.963 ± 0.003**	0.853 ± 0.005	0.936 ± 0.002
CBR-LargeTall	8.5 M	0.887 ± 0.004	0.957 ± 0.003	0.862 ± 0.003	0.936 ± 0.002
ResNet-18	11.5 M	**0.896 ± 0.004**	**0.964 ± 0.002**	0.881 ± 0.006	0.964 ± 0.001

Acknowledgements

Research was funded by the Priority Research Area Digiworld under the program Excellence Initiative – Research University at the Jagiellonian University in Kraków.

References

[1] A. Buades, B. Coll, and J.-M. Morel, Non-local means denoising, Image Processing On Line 1 (2011), 208-212.

[2] W. Bulten, G. Litjens, H. Pinckaers, P. Ström, M. Eklund, K. Kartasalo, M. Demkin, S. Dane, The PANDA challenge: Prostate cANcer graDe Assessment using the Gleason grading system (2020). https://doi.org/10.5281/zenodo.3715938.

[3] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, A simple framework for contrastive learning of visual representations, in: International conference on machine learning, PMLR, 2020, pp. 1597-1607.

[4] L. Jing and Y. Tian, Self-supervised visual feature learning with deep neural networks: A survey, IEEE Transactions on Pattern Analysis and Machine Intelligence (2020).

[5] M. Nauta, A. Jutte, J. Provoost, and C. Seifert, This Looks Like That, Because... Explaining Prototypes for Interpretable Image Recognition, arXiv preprint arXiv:2011.02863 (2020).

[6] M. Raghu, J. Gilmer, J. Yosinski, and J. Sohl-Dickstein, Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability, arXiv preprint arXiv:1706.05806 (2017).

[7] M. Raghu, C. Zhang, J. Kleinberg, and S. Bengio, Transfusion: Understanding transfer learning for medical imaging, arXiv preprint arXiv:1902.07208 (2019).

Address for correspondence

Corresponding author: Dawid Rymarczyk
E-mail: dawid.rymarczyk@student.uj.edu.pl