Evaluation of Contrastive Predictive Coding for Histopathology Applications

Karin Stacke KARIN.STACKE@LIU.SE
Department of Science and Technology (ITN), Linköping University, Sweden
Sectra AB, Linköping

Claes Lundström CLAES.LUNDSTROM@LIU.SE
Department of Science and Technology (ITN), Linköping University, Sweden
Center for Medical Image Science and Visualization (CMIV), Linköping University, Sweden
Sectra AB, Linköping

Jonas Unger JONAS.UNGER@LIU.SE
Gabriel Eilertsen GABRIEL.EILERTSEN@LIU.SE
Department of Science and Technology (ITN), Linköping University, Sweden
Center for Medical Image Science and Visualization (CMIV), Linköping University, Sweden

Editors: Emily Alsentzer[⊗], Matthew B. A. McDermott[⊗], Fabian Falck, Suproteem K. Sarkar, Subhrajit Roy[‡], Stephanie L. Hyland[‡]

Abstract

Recent advances in self-supervised learning for image data are closing the gap between unsupervised and supervised learning. However, the effectiveness of self-supervised methods has primarily been demonstrated for natural images. If the results would extrapolate to histopathology images, there could be significant benefits due to the reduced need for annotated data. In this paper, Contrastive Predictive Coding (CPC), one of the most promising state-of-the-art self-supervised methods, is extensively evaluated on histology data by varying a range of different parameters, including training objective, resolution, and data setup. From the results, we are able to draw important conclusions on the usefulness of CPC for digital pathology. We show strong evidence of the limitations of the learned representation for tumor classification, where only low-level information learned early during training, in the first CPC layers, is used. Furthermore, in our experiments, diversifying the distribution of the dataset (i.e., data from multiple organs or medical centers) does not lead to the model learning a more general representation. This study deepens the understanding of how the CPC model's objective relates to intrinsic characteristics of histology datasets and will help the development of effective self-supervised methods for histopathology.

Keywords: Histopathology, Self-supervised learning, Contrastive learning

1. Introduction

Recent advances in self-supervised methods have shown great success, closing the gap between supervised and unsupervised training for natural images (Oord et al., 2019; Chen et al., 2020a; Grill et al., 2020). Being able to harness this success for other types of image data is of great interest. Histopathology is one example, where the digitization of image slides has led to a wealth of digital data, but where annotated data is scarce. Being

able to utilize a pre-trained representation would greatly increase the potential for effective deep learning methods for histopathology applications.

One recently presented self-supervised learning method, Contrastive Predictive Coding (CPC), has been shown to be able to learn high-level features for multiple domains, such as vision, audio and reinforcement learning (Oord et al., 2019). By using a contrastive loss, the model learns to embed subsets of the input data in such a way that the representation within a single data point is kept similar. This forces the model to learn global, common features, rather than local variances. The general nature of this approach spurred our interest to evaluate it in our domain. In this paper, we present a rigorous evaluation of the CPC method for histopathology applications. A large number of configurations of the method are considered, multiple datasets are used for training, and the models are evaluated by linear tumor classification on three different tissue types.

The main contributions of this work are the conclusions that can be drawn from our large-scale study, summarized as follows:

- From investigating model specific parameters such as output dimensionality and training objective, as well as data specific parameters such as patch size and magnification, the results show that the best CPC configurations for histopathology applications are not the same as for natural images. Furthermore, the results show little correlation between the CPC objective and downstream classification performance.

- From layer-wise performance analysis, we find evidence that the CPC model is unable to learn high-level features relevant for tumor classification, and that only low-level features from the first layers (which are learned quickly) are used.

- Evaluation of different data distributions, varying tissue type(s), dataset size, and origin, show little difference in classification performance, even with widely different distributions.

2. Background and related work

Self-supervised learning (Sa, 1994) is a specialization of unsupervised learning, where a *pretext* task is formulated together with labels directly accessible from the data, such that the representations created for this task are useful for other, *downstream* tasks.

The increase in popularity of self-supervised learning for vision applications started with Doersch et al. (2015), continuing with Pathak et al. (2016); Zhang et al. (2016); Noroozi et al. (2017); Noroozi and Favaro (2016); Larsson et al. (2017), to mention a few. Recently, methods such as BiGAN (Donahue et al., 2017), CPC(v2) (Oord et al., 2019; Hénaff et al., 2019), MoCo (He et al., 2020), SimCLR(v2) (Chen et al., 2020a,b), and BYOL (Grill et al., 2020) have all shown high performance, quickly closing the gap between unsupervised and supervised learning. In this study, CPC was chosen as self-supervised method due to its general nature. The implementation is less coupled with the specific data type than methods which rely on augmentation strategies (such as SimCLR and BYOL). This allows us to compare the training setup and the learned representations across datasets, without large domain specific adaptations.

In the domain of histopathology, the number of presented self-supervised methods and applications are fewer. Gildenblat and Klaiman (2020) presented a method for self labeling based on spatial location of image patches. Using the assumption that spatially close patches are more similar than those farther away, they train a Siamese network with a contrastive loss such that this spatial re-

lationship is kept in the embedding space as well. Recently, Koohbanani et al. (2020) presented a self-supervised method tailored for histology data by combining multiple pretext task. They show that using the representations from the pretexts, they are able to significantly reduce the amount of annotated data needed for downstream tissue classification.

Despite the growing success of self-supervised methods, the understanding of how and why these methods work is limited. For example, Kolesnikov et al. (2019) showed that the success of self-supervised methods are as much related to the model architecture as to the pretext task. Goyal et al. (2019) presented results using millions of images, while Asano et al. (2020) showed that low-level features can be learnt from a single image. With this paper, we hope to expand the understanding of training self-supervised models on histopathology data.

Contrastive Predictive Coding (CPC) was presented by Oord et al. (2019) as a general self-supervised method for multiple types of high-dimensional data. The method is less coupled to specific datasets, compared to methods depending on augmentations (e.g., SimCLR (Chen et al., 2020a)), which may require significant effort to adapt to new domains (Grill et al., 2020). By using a contrastive loss, high-dimensional representations of subsets of each data point are used for predicting the *future* subsets of the same sample. For image data, this is constructed by dividing each input image into a 7x7 grid of overlapping crops. For each crop, the model predicts which of the representations belong to the 1–k crops directly below the target crop, by contrasting against N negative samples. The negative samples are taken as crops of images from the mini-batch. This way, representations from the same image are forced to be kept similar and encode more global features, instead of local variations (typically noise). We refer to Oord et al. (2019) for a more detailed description of the method.

3. Experimental Setup

In evaluating CPC for histology data, we focus on two separate aspects of the CPC training and its evaluation. The first investigates how the configuration of the CPC objective and optimization affects the downstream classification task, while the second focus on the data fed to the CPC. By considering different types of training and test data we are able to compare how the learned representation transfers between different domains (natural images, different tissue types), and how changing the distribution of the histology datasets affects the learned representation.

3.1. Evaluation

We evaluate the quality of the CPC learned representations by training a (supervised) linear logistic classifier on top of the frozen weights of the CPC model. As discussed by Kolesnikov et al. (2019), linear classification (compared to fine-tuning) is sufficient to give a strong indication of representational strength, in the context of a downstream task. During training, a snapshot of the model was extracted every 10th epoch for evaluation of the training progress. For more implementation details, we refer to supplementary material S1.

3.2. Contrastive Predictive Coding

In this study, we follow the CPC model implementation as in Löwe et al. (2019) [1]. This implementation differs slightly from the original implementation by Oord et al. (2019) as it does not use an auto-regression model.

1. Code available at https://github.com/loeweX/Greedy_InfoMax.

Table 1: CPC configuration parameters evaluated. Bold values indicate default values, underlines values indicate the optimally found value (see Section 4 from details).

Parameter	Value
Output dim.	32 128 512 **1024**
Prediction dir.	**1** 4
Patch size	**64x64** 128x128
Crop augm.	True **False**

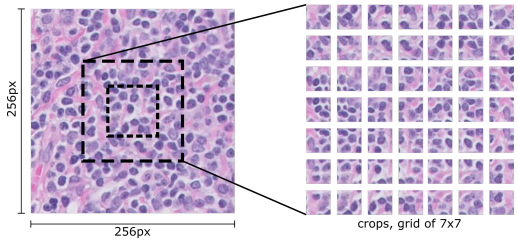

Figure 1: Example image, extracted at 20x magnification. The smaller square denote 64x64px patches, the larger 128x128px patches, from where a grid of 7x7 crops is exemplified.

This simplifies the model without loss of performance. The model was trained in an end-to-end fashion. The number of negative samples N was set to 16, and the number of predictions steps k to 5. Each image was cropped to 64x64 px (randomly during training, center crop during test), augmented with random flip, and color variation of the hue channel. Batch normalization was not used (as in Oord et al. (2019); Löwe et al. (2019); Hénaff et al. (2019)).

Hénaff et al. (2019) proposed a number of modifications of the original CPC-model configuration for image data. They found that increasing model and patch size, using an extended objective with multiple prediction directions and additional data augmentation help create better representations. In our experiments, we evaluate a number of the suggested additions/modifications, as shown in Table 1. In addition to parameter evaluation, the effect of the dataset was evaluated. Detailed description of the datasets is given below. From the histology datasets, patches were extracted at 20x magnification with patch size 256x256 pixels. An example image is shown in Figure 1, where patch sizes of 64x64px (smaller square) and 128x128px (larger square) are shown. All experiments ran on 4 Nvidia Tesla V100 GPUs, with a total time of approximately 2500 GPU hours.

3.3. Datasets

The data used in the experiments originate from four different datasets:

STL-10 (Coates et al., 2011): a dataset developed for unsupervised learning with a separate training set of unlabeled data, and a labeled test set, consisting of images from the ImageNet (Deng et al., 2009) dataset. Each image is of size 96x96 pixels.

CAMELYON17 (Litjens et al., 2018): 500 Haemotoxylin & Eosin (H&E) stained whole slide images (WSI's) of breast lymphnode tissue. This data was separated in two training sets, one smaller and one larger, with a common test set. Out of the 500 WSI's, only 50 WSI's contain tumor tissue. Out of these, 10 WSI's (20%) were selected as test set, where patches were sampled in a supervised way – dense sampling in tumor areas for increased class balance between tumor and non-tumor tissue (tumor patch percentage 17%). The remaining 40 WSI's were selected as training set, denoted CAM17 (tumor patch percentage 7%). The extracted patches can be considered as semi-supervised sampling, as the slide label was known, but sampling of the individual slides was done without reference to tissue annotations. The larger training set is a superset of CAM17,

including samples from all remaining WSI's (tumor patch percentage 0.7%). Please see supplementary material Table S1 for details. Evaluation on CAM17 test set is done by 5-fold cross validation, were 8 WSI's were used for training and 2 for testing.

AIDA-LNCO (Maras et al., 2019): 402 H&E-stained WSI's from regional lymph node metastasis in colon adenocarcinoma, originating from 38 patients. From the WSI's, patches were sampled in a supervised way, from annotated tumor/normal regions. The dataset was split at patient level, such that WSI's from approximately 70% of the patients were chosen as unsupervised training set, corresponding to 74% of the WSI's (tumor percentage 47%). The supervised dataset was divided (at patient level) in one training and one test set, with 48 and 58 WSI's respectively (tumor percentage 25% and 60%). A subset of the supervised training set was used as validation set for the unsupervised learning. Please see supplementary material Table S1 for details.

AIDA-SKIN (Lindman et al., 2019): 106 H&E-stained WSI's of skin tissue, from 71 patients. The dataset was split at patient level, where 80% were used for unsupervised training, and 20% for supervised training and test. For supervised evaluation, a 5-fold cross validation scheme was used, were WSI's from 12 patients were used for training and 3 for testing (1-2 WSI's per patient). The patches were sampled in a supervised way, where tumor and normal patches were sampled from annotated regions with corresponding labels. The unsupervised dataset consisted of 7% tumor patches, the supervised of 22%. Please see supplementary material Table S1 for details.

MIXED For training the CPC model, a dataset of three tissue types was constructed by combining the unsupervised training sets for CAM17, AIDA-LNCO, and AIDA-SKIN.

4. Training Contrastive Predictive Coding on Histology Data

In this section, experiments and results are presented. First, the impact of CPC configuration on the downstream classification accuracy is shown, followed by a presentation of the impact of the training dataset on the model representations.

4.1. Configuration of CPC

We investigated different configurations of the CPC method, by training and evaluating using the CAM17 dataset.

The CPC objective does not correlate with downstream classification performance. The first experiment evaluated the effect of training time and number of iterations. By evaluating the performance at every 10th epoch, we saw that the performance did not increase with increased training time, see Figure 2 (dark blue curve). After 10 epochs the performance was more or less stationary, i.e., the CPC model objective did *not* correlate with the downstream performance (the CPC loss kept decreasing, see Figure 3). This indicates that the contrastive learning objective is not helpful in tumor classification. We found that these results are consistent over multiple configurations and datasets. Random CPC (untrained) did however result in significantly lower performance (see Figure 2, black dotted lines), showing that the model indeed learns something useful, and that this happens very quickly. For succeeding experiments, we used 50 epochs as this point coincided with where the CPC training loss plateaus (Figure 3).

Tumor classification relies on low-level features. In order to more closely examine the reason behind the missing correlation between the CPC objective (where the loss keeps decreasing) and tumor classifica-

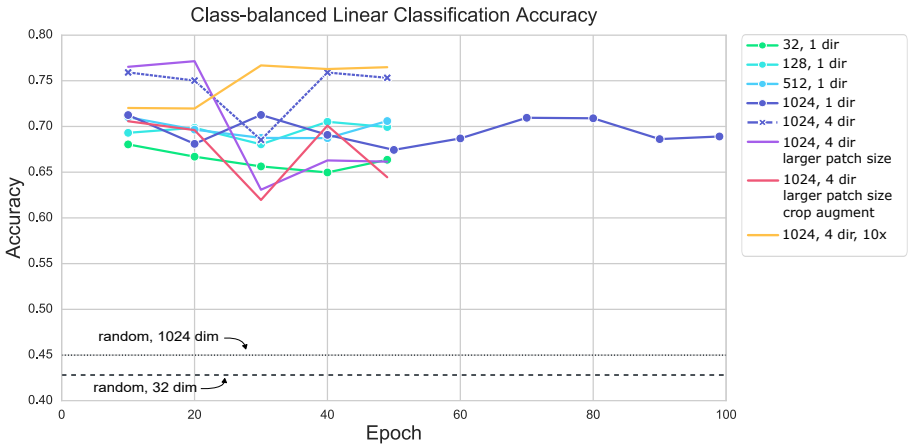

Figure 2: Accuracy over epochs for multiple configurations of the CPC method. All results are the class-balanced linear classification accuracy on the CAM17 test set. Random (untrained) CPC models with output dimension 32 and 1024 respectively are shown in black, dashed lines.

tion (which remains plateaued), we evaluated the representation at lower layers. Figure 4 (blue line) shows the linear classification accuracy per layer of a CPC model trained for 50 epochs on CAM17 data. The best separation between tumor and non-tumor tissue occurs already after the second block, i.e., very early in the model. This was different from a model trained and evaluated on STL-10 data (red line), where the highest accuracy is at the end. As high-level features are learned in deeper layers, these results give strong evidence that only low-level CPC features are relevant for tumor classification. This finding explains why the performance was not increasing with longer training, as earlier layers converge first (see supplementary material Figure S1 and Raghu et al. (2017)).

Direction invariant learning objective boosts performance. In the original implementation (Oord et al., 2019), the objective was only evaluated as prediction ability on 1–k steps *below* the target crop. As in Hénaff et al. (2019), adding more directions (up, left and right) indeed increased the classification accuracy (see Figure 2, dark blue dotted line). As WSI's are rotation invariant, the specific direction is of little interest, and adding more directions increases the de facto learning objective four times.

Larger field-of-view helps classification performance. Using the results from above, we kept the increased objective (with four prediction directions), and investigated how image content (size and scope) affect the representation. The image content of the field-of-view is determined by the magnification level and patch size. Hénaff et al. (2019) showed that a larger patch size improves performance for natural images. However, increasing the patch size in this setup from 64x64px to 128x128px at 20x magnification did not improve performance (Figure 2, purple line). Interestingly, keeping the patch size of 64x64px but reducing the magnification to 10x - which results in the same field-of-view but at different resolution - gave

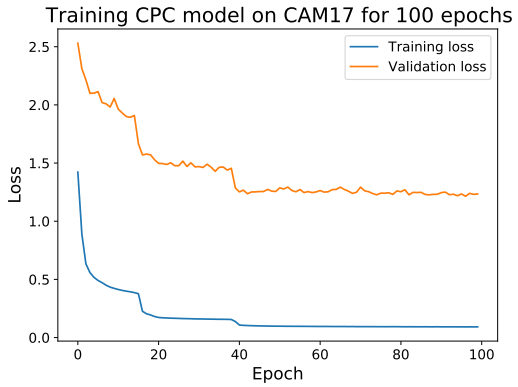

Figure 3: Loss of CPC model trained on CAM17. The loss plateaus after around 50 epochs.

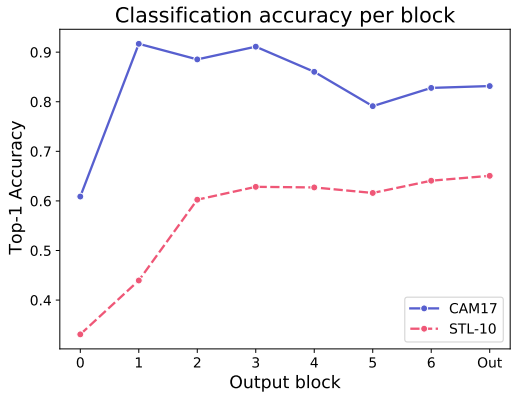

Figure 4: Linear tumor classification accuracy per block. Blue line is trained and evaluated on CAM17, red, dashed line is trained and evaluated on STL-10.

a considerable performance boost (Figure 2, yellow line).

Larger model size not needed for tumor classification. Hénaff et al. (2019) showed that increasing the model size and output dimension size from 1024 to 4096 led to better representations for object detection in natural images. In tumor classification however, the inter-class variance of different tissues can be considered much smaller than the inter-class variance between the classes in the STL-10 dataset (e.g., between airplane and dog). For histopathology, we saw that decreasing the output dimensionality from 1024, to 512 and 128 gave little performance difference. Further reducing to 32 indeed reduced performance. Together with the findings in the previous section, this indicates that the model is unable to extract high-level features relevant for tumor classification, and instead relies more on low-level features. This indicates that the current model size is sufficient, but the current training setup does not fully utilize its capacity for tumor classification. We believe that larger models or higher output dimensionality is unlikely to boost performance. Verification of this hypothesis is left for future work.

Individual crop augmentation does not improve classification performance. In Hénaff et al. (2019), individual crop augmentations was presented as a way of increasing performance. This is in some aspect similar to other recently presented self-supervised methods, such as Chen et al. (2020a); Grill et al. (2020), where the objective is to keep augmented versions of the same data sample close. In our case, by individually augmenting crops of the image sample, the model is forced to keep representations of the same image close/similar, invariant to the applied transformations. For this experiment, individual crops as well as the larger image patch were augmented with color jitter and random horizontal flip. The results however did not show any increase in performance (Figure 2, red line). As discussed in Grill et al. (2020), augmentations

for keeping representations close are specific to the type of data used. It is possible that a more extensive optimization of which augmentations to use could help boost performance.

4.2. Dataset impact

Based on the results above, we fix the CPC configuration to use 10x magnification, 64x64px patches, and the extended objective with four prediction directions. In this section, we evaluate the effect of the data used for training. All models are evaluated on the three test sets from CAM17, AIDA-LNCO and AIDA-SKIN, giving an indication as to how the learned representations generalize between tissues. The results are shown in Table 2, reported as mean linear classification accuracy training progress, taken from five epochs (as in Figure 2).

Tumor classification can use representations learned from natural images. By comparing CPC models trained on the STL-10 dataset and CAM17, we get an understanding on how domain specific the representation learned from CAM17 is. This comparison is shown in the top two rows of Table 2. As the results are in similar in terms of accuracy between the two CPC models, we can conclude that the representation learned from natural images in the STL-10 dataset generalizes to histopathology applications. There is no clear performance gain on training on CAM17, despite this being domain specific.

Increasing tissue variation of training data does not create more general representations. The CAM17 dataset contains only one tissue type, originating from five different medical centers (Litjens et al., 2018). Even if the size of the dataset (in terms of extracted patches) is larger than the STL-10 training set, the data may be too homogeneous for the model to be able to extract a general representation. To investigate if increased diversity is necessary for the CPC model, we train a model on the MIXED dataset (containing image patches from three different tissue types). As seen in Table 2, this did not result in a significant boost of performance. For AIDA-LNCO, the performance even dropped compared to STL-10 and CAM17 CPC models, even if the MIXED dataset contained images from the same distribution as the test set. Naively mixing data from different domains does not automatically generate a boost in performance, which was also observed in Feng et al. (2019). For visualizations of the representations, see supplementary material Section S3).

Adding more data from same distribution boosts performance. Continuing, the effect of increased dataset size from one domain only was investigated. The Full-CAM17 dataset contains patches sampled from all slides in the CAMELYON17 dataset, giving approximately 10x more data as compared to the CAM17 data. The added image patches were, however, all from normal tissue; no new tumor data was added. The downstream classification performance of this CPC model did increased on the CAM17 dataset, but was unchanged for other datasets (Table 2, last row). Despite reduced tumor percentage in the training set, the model was able to extract better features for tumor discrimination on same-distribution data, even if this improvement did not generalize to other tissue types. How much this boost is attributed to the increased training iterations vs increased data distribution is left for future work, but it is possible that a similar effect could be achieved with intensive data augmentation of a smaller training set (Asano et al., 2020).

Table 2: Mean linear classification accuracy (%) across five epochs (standard deviation in parenthesis), comparing different (frozen) CPC representations, trained on different datasets.

Unsupervised training set	Supervised test		
	CAM17	*AIDA-LNCO*	*AIDA-SKIN*
STL-10	82.3 (6.1)	79.7 (2.2)	81.2 (0.7)
CAM17	74.4 (2.7)	80.5 (1.7)	81.4 (1.7)
MIXED	79.0 (4.1)	78.0 (2.6)	83.0 (0.6)
FullCAM17	85.4 (5.0)	80.6 (3.7)	81.8 (0.3)

5. Discussion

Being able to learn relevant, general representations for downstream histopathology applications without the need for annotated data is highly desirable. From extensive experimentation, we have shown how CPC in its current formulation does not fully achieve this goal. The results in Figures 2 and 4 give strong evidence that only low-level features, learned very quickly, are discriminative for tissue classification. This is consistent with the result that a CPC model trained on STL-10 data have similar performance compared to CPC models trained on histology data. The model is able to extract these features from many different types of data. According to Asano et al. (2020), this could potentially be done from one single image.

Thus, the image understanding that is required to perform the CPC task on histology data does not require features which are relevant for tumor tissue discrimination. This is in contrast to the features learned for natural images. An important difference between the datasets are that the task of separating natural image classes (such as cars and planes) requires less fine-grained features than those required to separate tissue types in histology data. For pathology, low-level features are enough to achieve a *basic* separation between normal and tumor tissue, but for detecting the fine grained differences, required to reach high performance on pathology data, higher-level features are required. These are not learned by the CPC.

Another problem with pathology data is that domain differences, for example caused by different tissue type or scanning protocols at different medical centers, have strong differentiating characteristics. These characteristics have been shown to be encoded when training supervised learning models (Stacke et al., 2019; Lafarge et al., 2017). As nothing in the self-supervised learning objective is hindering the model from using these features, they are likely to be encoded in the CPC representation as well (see supplementary material S3 for UMAP visualizations of the representation which indicate this). Great care needs to be taken when training unsupervised and self-supervised systems in order such that these characteristics are not allowed to hamper the learning of other, to the downstream task, relevant features. A strategy of either removing or reducing extraneous features is needed, or a learning objective that forces the model to pay little attention to them.

Although the evaluation has focused on tumor classification as the downstream task, the preceding reasoning also extends to other tasks. Thus, we believe that general conclu-

sions can be drawn from the results, which are valid for other downstream tasks.

6. Conclusion

We have presented a rigorous evaluation of CPC on histopathology data. Multiple configurations and datasets have been used for training CPC models, and evaluation has been performed on multiple datasets of different tissue types. We have presented strong evidence that only low-level features, learned very early in the training process, are useful for downstream tumor classification. We have shown that these representations can be learned from both histology and natural images. Furthermore, naively diversifying the distribution of a histopathology dataset by adding data from different medical centers, scanners or tissues will primarily introduce extraneous features which impede the model from learning relevant features for downstream applications.

Acknowledgments

This work was supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP-AI), the research environment ELLIIT, and VINNOVA, grant 2017-02447 (AIDA).

References

YM Asano, C Rupprecht, and A Vedaldi. A critical analysis of self-supervision, or what we can learn from a single image. In *International Conference on Learning Representations*, 2020.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A Simple Framework for Contrastive Learning of Visual Representations. *arXiv:2002.05709 [cs, stat]*, 2020a.

Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. Big Self-Supervised Models are Strong Semi-Supervised Learners. *arXiv:2006.10029 [cs, stat]*, 2020b.

Adam Coates, Honglak Lee, and Andrew Y Ng. An Analysis of Single-Layer Networks in Unsupervised Feature Learning. *AISTATS*, page 9, 2011.

J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

Carl Doersch, Abhinav Gupta, and Alexei A. Efros. Unsupervised Visual Representation Learning by Context Prediction. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1422–1430, 2015.

Jeff Donahue, Philipp Krähenbühl, and Trevor Darrell. Adversarial Feature Learning. In *5th International Conference on Learning Representations, ICLR 2017, Conference Track Proceedings*. OpenReview.net, 2017.

Zeyu Feng, Chang Xu, and Dacheng Tao. Self-Supervised Representation Learning From Multi-Domain Data. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3244–3254, Seoul, Korea (South), 2019. IEEE. ISBN 978-1-72814-803-8.

Jacob Gildenblat and Eldad Klaiman. Self-Supervised Similarity Learning for Digital Pathology. *arXiv:1905.08139 [cs]*, 2020.

Priya Goyal, Dhruv Mahajan, Abhinav Gupta, and Ishan Misra. Scaling and benchmarking self-supervised visual representation learning. In *2019 IEEE/CVF International Conference on Computer Vi-*

sion, ICCV 2019, pages 6390–6399. IEEE, 2019.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap Your Own Latent: A New Approach to Self-Supervised Learning. *arXiv:2006.07733 [cs, stat]*, 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, Las Vegas, NV, USA, 2016. IEEE. ISBN 978-1-4673-8851-1.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum Contrast for Unsupervised Visual Representation Learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9726–9735, 2020.

Olivier J. Hénaff, Aravind Srinivas, Jeffrey De Fauw, Ali Razavi, Carl Doersch, S. M. Ali Eslami, and Aaron van den Oord. Data-Efficient Image Recognition with Contrastive Predictive Coding. *arXiv:1905.09272 [cs]*, 2019.

Alexander Kolesnikov, Xiaohua Zhai, and Lucas Beyer. Revisiting self-supervised visual representation learning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2019*, pages 1920–1929. Computer Vision Foundation / IEEE, 2019.

Navid Alemi Koohbanani, Balagopal Unnikrishnan, Syed Ali Khurram, Pavitra Krishnaswamy, and Nasir Rajpoot. SelfPath: Self-supervision for Classification of Pathology Images with Limited Annotations. *arXiv:2008.05571 [cs, eess]*, 2020.

Maxime W. Lafarge, Josien P. W. Pluim, Koen A. J. Eppenhof, Pim Moeskops, and Mitko Veta. Domain-Adversarial Neural Networks to Address the Appearance Variability of Histopathology Images. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, Lecture Notes in Computer Science. Springer International Publishing, 2017.

Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Colorization as a Proxy Task for Visual Understanding. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 840–849, 2017.

Karin Lindman, Jerónimo F. Rose, Martin Lindvall, and Caroline Bivik Stadler. Skin data from the visual sweden project droid, 2019. URL doi:10.23698/aida/drsk.

Geert Litjens, Peter Bandi, Babak Ehteshami Bejnordi, Oscar Geessink, Maschenka Balkenhol, Peter Bult, Altuna Halilovic, Meyke Hermsen, Rob van de Loo, Rob Vogels, Quirine F. Manson, Nikolas Stathonikos, Alexi Baidoshvili, Paul van Diest, Carla Wauters, Marcory van Dijk, and Jeroen van der Laak. 1399 H&E-stained sentinel lymph node sections of breast cancer patients: the CAMELYON dataset. *GigaScience*, 7(6), 2018.

Sindy Löwe, Peter O'Connor, and Bastiaan S. Veeling. Putting an end to end-to-end: Gradient-isolated learning of representations. In *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, pages 3033–3045, 2019.

Gordon Maras, Martin Lindvall, and Claes Lundstrom. Regional lymph node metastasis in colon adenocarcinoma, 2019. URL doi:10.23698/aida/lnco.

Leland McInnes, John Healy, and James Melville. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv:1802.03426 [cs, stat]*, December 2018.

Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *Computer Vision - ECCV, 14th European Conference, 2016, Proceedings, Part VI*, volume 9910 of *Lecture Notes in Computer Science*, pages 69–84. Springer, 2016.

Mehdi Noroozi, Hamed Pirsiavash, and Paolo Favaro. Representation learning by learning to count. In *IEEE International Conference on Computer Vision, ICCV 2017*, pages 5899–5907. IEEE Computer Society, 2017.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation Learning with Contrastive Predictive Coding. *arXiv:1807.03748 [cs, stat]*, 2019.

Deepak Pathak, Philipp Krähenbühl, Jeff Donahue, Trevor Darrell, and Alexei A. Efros. Context Encoders: Feature Learning by Inpainting. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2536–2544, 2016.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. SVCCA: singular vector canonical correlation analysis for deep learning dynamics and interpretability. In *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)*, pages 6076–6085, 2017.

Virginia R. de Sa. Learning Classification with Unlabeled Data. In *Advances in Neural Information Processing Systems 6 (NeurIPS 1994)*, pages 112–119. Morgan-Kaufmann, 1994.

Karin Stacke, Gabriel Eilertsen, Jonas Unger, and Claes Lundström. A Closer Look at Domain Shift for Deep Learning in Histopathology. *arXiv:1909.11575 [cs]*, September 2019.

Richard Zhang, Phillip Isola, and Alexei A. Efros. Colorful image colorization. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III*, volume 9907 of *Lecture Notes in Computer Science*, pages 649–666. Springer, 2016.

Supplementary Material

Below follow supplementary material for the article Evaluation of Contrastive Predictive Coding for Histopathology Applications.

S1. Implementation Details

PyTorch was used for all experiments. The CPC model was trained using the ResNet-v2 architecture (He et al., 2016), with the implementations from Löwe et al. (2019), on 4 Nvidia Tesla V100 GPUs, each batch size 32. Adam optimizer was used with learning rate of 2e−4. For evaluation, we append a linear logistic classifier on the frozen CPC model weights. The linear layer was for histology datasets trained for 10 epoch, for STL-10 test set for 50 epochs. Adam optimizer with learning rate of 5e−4 was used. The dataset sizes, in terms of number of whole slide images and extracted patches, are listen in Table S1.

S2. Learning dynamics

Figure S1 shows the learning dynamics of the CPC model trained on CAM17 data, showing bottom up training where the lower layers learn first.

S3. UMAP visualizations

Figure S2 and S3 show the UMAP (McInnes et al., 2018) visualizations of the embeddings from two CPC models, trained on STL-10 and MIXED data respectively. Both embeddings show data from the MIXED dataset, where the each input patch are drawn at its corresponding position. The left embedding in each figure shows the output from the first (i.e., shallow) layer block, the right are the embedding from the output layer. There is a clearer clustering when training on MIXED data, where same tissue types (such as fat tissue) are separated in different clusters. The clusters are highly correlated with origin (not shown).

Table S1: Number of WSI's (# samples in parenthesis) in datasets. * indicate the dataset is sampled in a supervised way, † 5-fold cross validation. The total supervised CAM17 dataset consisted of 30280 patches. The total supervised AIDA-SKIN dataset consisted of 90093 patches.

	Unsupervised		*Supervised*	
	Training	*Validation*	*Training*	*Test*
CAM17	36 (209529)	4 (14537)	8*†	2*†
Full CAM17	441 (2043958)	49 (227394)		
AIDA-COLON	296* (200445)	35* (12840)	48* (46876)	58* (43282)
AIDA-SKIN	61* (152128)	8* (34972)	22*†	5*†

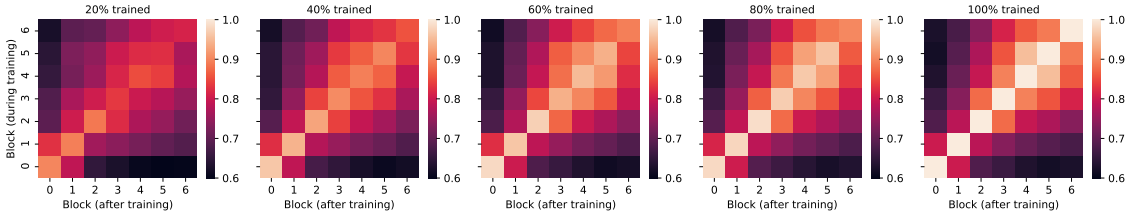

Figure S1: SVCCA similarities showing the learning dynamics of the CPC model trained on CAM17 dataset.

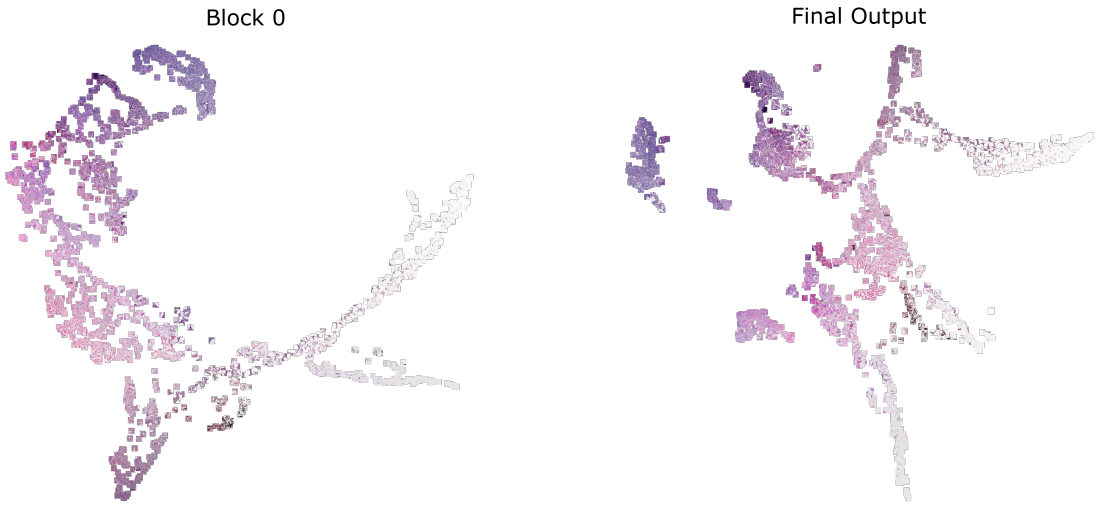

Figure S2: UMAP visualization of a CPC model trained on STL-10, output from the first block (left) and final output (right) with data from the MIXED dataset, where input images are shown at their corresponding position of the UMAP embedding.

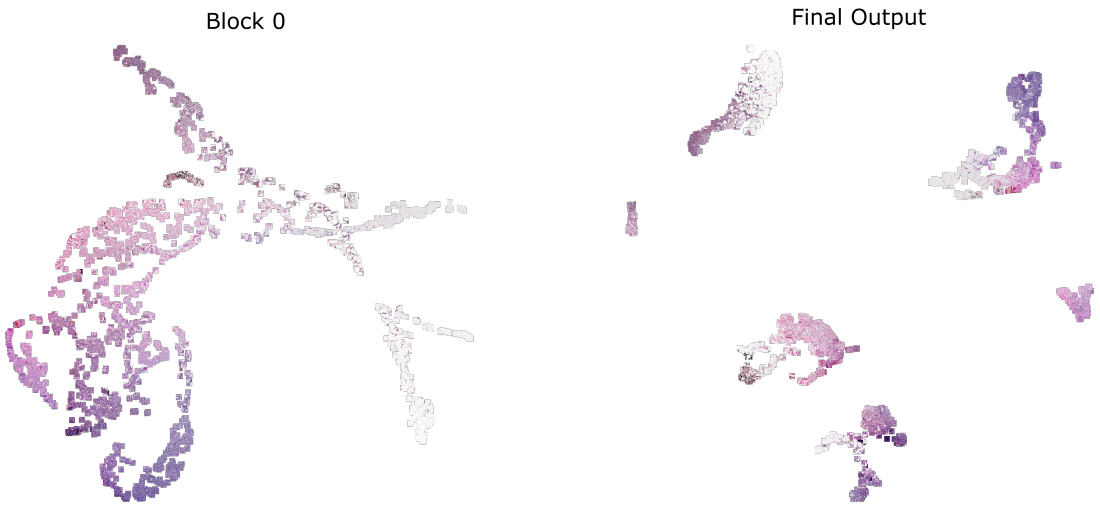

Figure S3: UMAP visualization of a CPC model trained on the MIXED dataset, output from the first block (left) and final output (right) with data from the MIXED dataset, where input images are shown at their corresponding position of the UMAP embedding.