Rosalie Kletzander*, Petr Kuritcyn, Volker Bruns, Markus Eckstein, Carol Geppert, Arndt Hartmann, and Michaela Benz

Domain Transfer in Histopathology using Multi-ProtoNets with Interactive Prototype Adaptation

https://doi.org/10.1515/cdbme-2023-1123

Abstract: Few-shot learning addresses the problem of classification when little data or few labels are available. This is especially relevant in histopathology, where labeling must be carried out by highly trained medical experts. Prototypical Networks promise transferability to new domains by using a pre-trained encoder and classifying by way of a prototypical representation of each class learned with few samples. We examine the applicability of this approach by attempting domain transfer from colon tissue (for training the encoder) to urothelial tissue. Furthermore, we address the problems arising from representing a class via a small amount of representatives (prototypes) by testing two different prototype calculation strategies. We compare the original "Prototype per Class" (PPC) approach to our "Prototype per Annotation" (PPA) method, which calculates one prototype for each example annotation made by the pathologist. We test the domain transfer capability of our approach on a dataset of 55 whole slide images (WSIs) containing six subtypes of urothelial carcinoma in two granularities: "Superclasses", which combines the tumorous subtypes into a single "tumor" class on top of a aggregated "healthy" and additional "necrosis" class, and "subtypes", which considers all eleven classes separately. We evaluate the classic PPC approach as well as our PPA approach on this data set. Our results show that the adaptation of the Prototypical Network from colon tissue to urothelial tissue was successful, yielding an F1 score of 0.91 for the "superclasses". Furthermore, the PPA approach performs very comparably to the PPC strategy. This makes it a viable alternative that places more value on the intent of the pathologist during annotation.

Keywords: few-shot learning, Prototypical Networks, histopathology, urothelial carcinoma, domain transfer, prototype calculation strategy

***Corresponding author: Rosalie Kletzander,** Fraunhofer Institute for Integrated Circuits IIS, Erlangen, Germany, e-mail: rosalie.kletzander@iis.fraunhofer.de
Petr Kuritcyn, Volker Bruns, Michaela Benz, Fraunhofer Institute for Integrated Circuits IIS, Erlangen, Germany
Markus Eckstein, Carol Geppert, Arndt Hartmann, Institute of Pathology, University Hospital Erlangen-Nürnberg, Erlangen, Germany

1 Introduction

In the last decade, neural networks have established themselves as the most popular method for image classification. Large amounts of labeled data used to train these neural networks have made extremely high classification accuracy possible in many areas of application. However, sufficient amounts of labeled data are not always available, e.g., due to the rarity of the class, or the high cost of labeling. This is especially relevant in histopathology, where labeling is dependent on highly trained medical experts. To enable accurate classification in cases such as these where only small amounts of labeled data is accessible, the field of few-shot learning has emerged.

One few-shot learning approach based on parametrized models is Prototypical Networks by Snell et al. [6]. Prototypical Networks learn an encoding so that representations of each class will form a cluster in the feature space. The clusters are then represented by a single prototype, which is the mean of the samples ("supports") of the class. For classification, the encoding of a query image is compared to all of the prototypes and the class of the prototype with the smallest euclidean distance is assigned.

In the domain of histopathology, Prototypical Networks have the potential to enable very adaptable classifiers that can learn an encoding based on a data-rich use case and then use the flexibility of the design to easily transfer to a new domain by adding small amounts of supports in the new domain. Possible domain transfers include different stainings, scanners or organs. A challenge of the Prototypical Network approach, however, is the calculation of the prototypes, which can be influenced by outliers, or unstable due to an unrepresentative support set. Classic prototype calculation methods calculate the prototype(s) from the entire support set of each class. While this is the most straightforward approach, it omits semantic information pertaining to the supports, specifically groupings based on the relative position of supports to each other e.g., same annotation or same whole slide image (WSI).

This group-agnostic approach may not be very effective in an interactive setting where a pathologist adapts a Prototypical Network by adding annotations as needed, and checks the classification results after each new annotation. For example, if

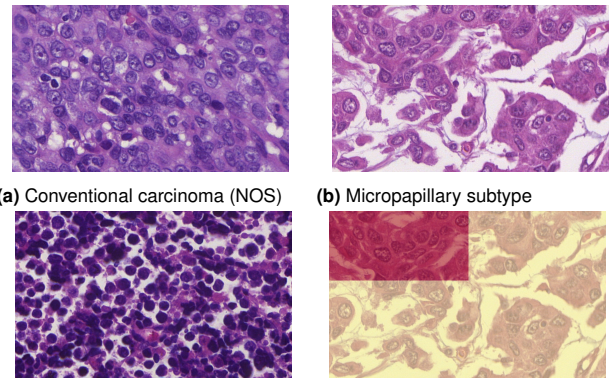
the pathologist sees an incorrectly classified area, they would annotate the area with the correct class and expect the classification results to respond accordingly. However, adding a single annotation (i.e. a limited number of supports) would most likely shift the associated prototype but may not have an effect on actual classification results, e.g., if the number of newly added supports is much smaller than the total number of supports, it will not have a large impact on the prototype, which is the mean value. This can be especially problematic when morphologically diverse classes are grouped together, such as when adapting a classifier to recognize healthy and tumorous tissue in cases where there are many tumor subtypes such as with the variant histological subtypes of urothelial carcinoma. Here, small annotations of rare subtypes must be recognized as tumor tissue as well as more common variants which may be represented by a much higher number in the support set.

In order to address this problem of unreactive adaptation, we propose an alternative prototype calculation method, which gives each annotation a greater impact on the classification results by calculating a prototype for each annotation instead of over all of the supports of a class. This effectively creates semantic subgroups within the supports that resemble the valuable knowledge applied by the pathologist by selecting a specific annotation. Furthermore, we test the general domain transfer ability of Prototypical Networks, specifically MultiProto-Nets [2], by using an encoder trained on colon tissue which is then adapted to urothelial tissue.

2 Related Work

Domain transfer using few-shot methods in medical imaging has been attempted in many different ways, e.g., by fine-tuning a pre-trained network, or by learning a feature space that can be translated to new tasks [3]. Domain transfer specifically with Prototypical Networks has been tested by Deuschel et al. [2] for various scanners on a single tissue type with a single staining.

Concerning prototype optimization, Snell et al. attempt to improve their prototypes by adding a semi-supervised element to their training step [5]. Liu et al. [4] propose a semi-supervised prototype rectification step applied after initial training in order to improve prototype robustness. Deuschel et al. [2] calculate multiple prototypes per class using k-means in order to represent classes with more heterogeneous clusters in the feature space.



(a) Conventional carcinoma (NOS) **(b)** Micropapillary subtype

(c) Neuroendocrine subtype **(d)** Erroneous classification

Fig. 1: a), b), c) show three different subtypes. d) shows the misclassification of micropapillary tissue (yellow) as NOS (red) in the top left corner. The tumor cells themselves look very similar, explaining the misclassification.

3 Materials and Methods

Our approach to testing the domain transferability of the Prototypical Network and comparing the results of the original Prototype per Class (PPC) to our proposed Prototype per Annotation (PPA) method, is split into two parts: First, the domain transferability of the Prototypical Network, specifically the Multi-ProtoNet [2] from colon to urothelial carcinoma is evaluated using the classic PPC approach. Then, to test the viability of the PPA approach, the same data is processed with the new prototype calculation strategy.

We use a dataset of hematoxylin and eosin (HE)-stained WSIs containing urothelial carcinomas in the resolution 0.194 $\mu\mathrm{m/px}$. Urothelial cancer lends itself well to our purpose for several reasons: It differentiates into a number of subtypes, some of which are morphologically very similar (conventional and micropapillary carcinoma), others of which are morphologically very different (e.g., neuroendocrine carcinoma), as shown in Figures 1a, 1b and 1c. This permits us to test our approach on a number of diverse classes. The subtypes can be summarized in a "tumor" superclass [1], which enables us to test our approach on different granularities of classes, the "subtypes" and the "superclasses", where the descriptor "subtype" is used both for the subtypes of urothelial carcinoma, as well as for the "subtypes" of healthy tissue.

The dataset consists of 55 WSIs, where 18 are used for adaptation and 37 are used for testing. Table 1 shows the subtypes, superclasses and the number of tiles and slides used for adaptation and testing. Each annotation used in these experiments is split into tiles of 224x224 pixels (50x50μm^2), which are then used as input for the MultiProto-Net.

The encoder backbone of the Multi-ProtoNet is an EfficientNetB0 [7] neural network trained on the dataset of more

subtype	superclass	WSIs (tiles) for adaptation	WSIs (tiles) for testing
NOS	**tumor**	**3 (680)**	**18 (3637)**
neuroendocrine	**tumor**	**3 (439)**	**5 (1781)**
sarcomatoid	**tumor**	**3 (521)**	**4 (2006)**
plasmacytoid	**tumor**	**3 (352)**	**4 (1695)**
micropapillary	**tumor**	**3 (520)**	**2 (2509)**
squamous	**tumor**	**3 (598)**	**2 (2335)**
connective tissue	**healthy**	**3 (450)**	**27 (8421)**
fat	**healthy**	**3 (917)**	**21 (9064)**
muscle	**healthy**	**3 (520)**	**23 (4657)**
inflammation	**healthy**	**3 (104)**	**22 (597)**
necrosis	**necrosis**	**3 (314)**	**13 (2126)**

Tab. 1: Urothelial carcinoma dataset. "NOS" refers to "not otherwise specified", i.e. conventional urothelial carcinoma

than two million 224x224px image patches which was also used by Deuschel et al. [2]: These patches are extracted from 92 HE stained colon tissue sections from adenocarcinoma resections and are assigned to seven tissue classes. Using the EfficientNetB0 backbone trained on colon carcinoma to create a classifier for urothelial carcinoma is the basis of our experiment on domain adaptability for the Multi-ProtoNet approach.

3.1 Domain Transfer from Colon Tissue to Urothelial Tissue

We conduct our experiments in two different granularities, yielding the "superclasses" classifier and the "subtypes" classifier. Both classifiers are adapted using the same annotations, listed in Table 1, in one case with the superclass label and in the other with the subtype label. Each subtype class is represented by three annotations, each from a different WSI in order to prevent "overfitting" to a single WSI/patient.

For the "superclasses" classifier, this results in 18 annotations for the tumor class, twelve annotations for the healthy class, and three annotations for the necrosis class. As we are utilizing the Multi-ProtoNet approach, we allow multiple prototypes per class, specifically, we choose six prototypes per class, which are calculated using k-means. We choose six prototypes so that each of the six considered tumor subtypes in our dataset can be represented in feature space, even though they are all grouped together into a single class. The healthy and necrosis classes are also represented by six prototypes although they contain fewer classes, as the Multi-ProtoNet method sets the same number of prototypes for all classes.

For the "subtypes" classifier, we select three prototypes per subtype, as established in [2]. Coincidentally, this is also the number of annotations for each subtype.

Using these prototypes, we run the "subtypes" and "superclasses" classifiers on our test data. The test set contains at least two up to 27 WSIs per "subtype", as shown in Table 1. Each of the WSIs of the test set contains one or more ground truth annotations, which are used for comparison with the classification results.

3.2 Prototype Calculation Strategy

In the next step, we test our PPA calculation strategy. The setup is identical to the original PPC setup used in the domain adaptability experiment, there being two different classifiers adapted at the "subtype" and "superclass" granularities. As opposed to the PPC calculation strategy, where the prototypes are calculated with k-means using all of the supports of each class, the PPA strategy calculates a single average feature vector per annotation of each class. This yields 18 prototypes for tumor, 12 prototypes for healthy and 3 prototypes for necrosis in the "subtypes" classifier. The "superclasses" classifier also contains the same numbers of prototypes as the "subtypes" classifier. As a matter of fact, the prototypes are identical in both classifiers, as the annotations and therefore the supports do not change.

4 Results

The overall accuracy, average precision, average recall and average F1 scores of all four classifiers are listed in Table 2. The precision, recall and F1 scores are calculated for each class individually and then averaged over all the classes.

The "superclass" PPC classifier ("super PPC") achieves an accuracy of 93.6% on the test set of 38828 tiles, with an average F1 score of 0.912, as shown in Table 2. This is significantly higher than the values reached by the "subtype" PPC classifier ("sub PPC"), which differentiates between eleven classes instead of three, with 68.3% and 0.572, respectively. An investigation of the errors showed that the vast majority of errors introduced by the finer granularity were mixups between the tumor subtypes. We determined this by summing up the confusion matrices of the subtypes to yield a superclass confusion matrix, and calculated the scores on this combined confusion matrix, shown in the line "Σ sub PPC" of Table 2.

The PPA classifiers show very similar results, with an accuracy of 92.9% and an average F1 score of 0.913 for the "superclass" classifier ("super PPA") and 67.5% and 0.565, respectively for the "subtype" classifier ("sub PPA"). Here, the summed "sub PPA" confusion matrix scores are identical to

	accuracy	avg. prec.	avg. recall	avg. F1
super PPC	**93.6%**	**89.8%**	**93.0%**	**0.912**
sub PPC	**68.3%**	**56.6%**	**59.1%**	**0.572**
Σ sub PPC	**93.4%**	**90.3%**	**93.0%**	**0.916**
super PPA	**92.9%**	**90.8%**	**91.9%**	**0.913**
sub PPA	**67.5%**	**56.4%**	**58.1%**	**0.565**
Σ sub PPA	**92.9%**	**90.8%**	**91.9%**	**0.913**

Tab. 2: Results of the PPA and PPC metrics for super- and subclasses. The Σ sub PPx results are calculated from the subclass results combined by superclass i.e. the 3-class problem.

the "super PPA" scores, because the prototypes are equal for both classifiers, as explained above.

Altogether, the PPA classifiers performed very similarly to the PPC (i.e. original strategy) classifiers, with only slightly lower scores for the "subtypes" (0.007 lower F1 score) and mixed scores for the "superclasses" (0.001 higher F1 score, but lower recall).

5 Discussion and Conclusion

In general, the transfer from colon to urothelial tissue using only 33 annotations was successful, as shown by the high scores of the superclass classifiers. Even the significantly worse performance of the subtypes classifiers was mostly satisfactory, seeing as we were attempting to distinguish between eleven, partly very similar tissue types. We qualitatively evaluated extreme cases of mixups, e.g., more than 50% of the ground truth tumor subtype being misclassified as a different tumor subtype. In many of these cases, the classification results were reasonable, since the subtypes can contain very similar cells e.g., the tumor cells in Figure 1a and Figure 1b. In some of these cases, the correct classification seems unachievable by evaluating a tile individually, since correct classification would require further surrounding context.

Concerning the comparison of the two prototype calculation strategies, the similar results are promising for moving forward with the PPA method, since it most likely offers advantages over the PPC method for interactive classifier adaptation where it allows capturing the intention behind a carefully placed annotation. The PPA method will likely be superior at ensuring that "corrective" annotations will change a local false prediction to the desired one in the subsequent run - a problem that is not reflected in the standard quality metrics reported above. Furthermore, it is likely that the PPA method would outperform the PPC method in less favorable circumstances. In our experiments, we chose the parameters very advantageously for the PPC approach, e.g., by setting the number of prototypes to the exact number of tumor subtypes and gathering a balanced number of supports for all the classes. If the number of subtypes is not known, it is easy to set a too-low number of prototypes, which would presumably lead to many more misclassifications due to the lack of representation in the prototypes.

Going forward, we expect to continue to extensively experiment with our PPA calculation strategy, in order to optimize the interactive adaptation process, enabling pathologists to apply their expertise in a concise and effective way.

Author Statement

This work was supported by the Bavarian Ministry of Economic Affairs, Regional Development & Energy through the Center for Analytics - Data - Applications (ADA-Center) within "Bayern Digital II" and by the BMBF (16FMD01K, 16FMD02, 16FMD03). There were no conflicts of interest.

References

[1] Anna J. Black and Peter C. Black. Variant histology in bladder cancer: diagnostic and clinical implications. *Translational Cancer Research*, 9(10), 2020. ISSN 2219-6803. URL https://tcr.amegroups.com/article/view/41634.

[2] Jessica Deuschel, Daniel Firmbach, Carol I. Geppert, Markus Eckstein, Arndt Hartmann, Volker Bruns, Petr Kuritcyn, Jakob Dexl, David Hartmann, Dominik Perrin, Thomas Wittenberg, and Michaela Benz. Multi-prototype few-shot learning in histopathology. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 620–628, 10 2021.

[3] Hao Guan and Mingxia Liu. Domain adaptation for medical image analysis: A survey. *CoRR*, abs/2102.09508, 2021. URL https://arxiv.org/abs/2102.09508.

[4] Jinlu Liu, Liang Song, and Yongqiang Qin. Prototype rectification for few-shot learning. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 741–756, Cham, 2020. Springer International Publishing. ISBN 978-3-030-58452-8.

[5] Mengye Ren, Eleni Triantafillou, Sachin Ravi, Jake Snell, Kevin Swersky, Joshua B. Tenenbaum, Hugo Larochelle, and Richard S. Zemel. Meta-learning for semi-supervised few-shot classification. *CoRR*, abs/1803.00676, 2018. URL http://arxiv.org/abs/1803.00676.

[6] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/cb8da6767461f2812ae4290eac7cbc42-Paper.pdf.

[7] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. *CoRR*, abs/1905.11946, 2019. URL http://arxiv.org/abs/1905.11946.