



# Interpretable Medical Image Classification Using Prototype Learning and Privileged Information

Luisa Gallée<sup>1</sup> , Meinrad Beer<sup>2</sup> , and Michael Götz<sup>1,3</sup>  (✉)

<sup>1</sup> Experimental Radiology, University Hospital Ulm, Ulm, Germany  
{luisa.gallee,michael.goetz}@uni-ulm.de

<sup>2</sup> Department of Diagnostic and Interventional Radiology,  
University Hospital Ulm, Ulm, Germany

<sup>3</sup> i2SouI - Innovative Imaging in Surgical Oncology Ulm,  
University Hospital Ulm, Ulm, Germany

**Abstract.** Interpretability is often an essential requirement in medical imaging. Advanced deep learning methods are required to address this need for explainability and high performance. In this work, we investigate whether additional information available during the training process can be used to create an understandable and powerful model. We propose an innovative solution called *Proto-Caps* that leverages the benefits of capsule networks, prototype learning and the use of privileged information. Evaluating the proposed solution on the LIDC-IDRI dataset shows that it combines increased interpretability with above state-of-the-art prediction performance. Compared to the explainable baseline model, our method achieves more than 6% higher accuracy in predicting both malignancy (93.0%) and mean characteristic features of lung nodules. Simultaneously, the model provides case-based reasoning with prototype representations that allow visual validation of radiologist-defined attributes.

**Keywords:** Explainable AI · Capsule Network · Prototype Learning

## 1 Introduction

Deep learning-based systems show remarkable predictive performance in many computer vision tasks, including medical image analysis, and are often comparable to human performance. However, the complexity of this technique makes it challenging to extract model knowledge and understand model decisions. This limitation is being addressed by the field of Explainable AI, in which significant progress has been made in recent years. An important line of research is the use of inherently explainable models, which circumvent the need for indirect, error-prone on-top explanations [14]. A common misconception is that the additional explanation comes with a decrease in performance. However, Rudin et al. [14]

and others have already pointed out that this can be avoided by designing algorithms that build explainability into the core concept, rather than just adding it on top. Our work proves this once again by providing a powerful and explainable solution for medical image classification.

A promising approach for interpretability is the use of **Privileged Information**, i.e. information that is only available during training [19,20]. Besides using the additional knowledge to improve performance, it can also help to increase explainability, as has already been shown using the LIDC-IDRI dataset [3]. In addition to the malignancy of the lung nodules, which is the main goal of the prediction task, the radiologists also marked certain nodule characteristics such as sphericity, margin or spiculation. Shen et al. [16] used the attributes with a hierarchical 3D CNN approach, demonstrating the potential of using this privileged information. LaLonde et al. [11] extended this idea using capsule networks, a technique for learning individual, encapsulated representations rather than general convolutional layers [1,15]. This method was used to jointly learn the predefined attributes in the capsules and their associations with the classification target, i.e. malignancy. Explainability is enabled by providing additional attribute values that are essential to the model output. However, the predicted, possibly incorrect scores for the attributes must be trusted, which raises the question of whether there is a way to validate the predictions.

**Prototype Networks** are another line of research implementing the idea that the representations of images cluster around a prototypical representation for each class [17]. The goal is to find embedded prototypes (i.e. examples) that best separate the images by their classes [5]. This idea has been applied to various methods, such as unsupervised learning [13], few- and zero-shot learning [17,18,22], as well as for capsule networks [21], however without the use of privileged information. A successful approach is prototypical models with case-based reasoning, which justify their prediction by showing prototypical training examples similar to the input instance [4,12]. This idea can be used for region-wise prototypical samples [6]. However, these networks can only tell which prototypical samples resemble the query image, not why. Similar to attention models, regional explanations are learned and provided [23,24]. It is up to the user to guess which features of the image regions are relevant to the network and are exemplified by the prototypes.

Our method addresses the limitations of privileged information-based and prototype-based explanation by combining case-based visual reasoning through exemplary representation of high-level attributes to achieve explainability and high-performance. The proposed method is an image classifier that satisfies explainable-by-design with two elements: First, decisive intermediate results of a high-performance CNN are trained on human-defined attributes which are being predicted during application. Second, the model provides prototypical natural images to validate the attribute prediction. In addition to the enhanced explainability offered by the proposed approach, to our knowledge the proposed method outperforms existing studies on the LIDC-IDRI dataset.

The main contributions of our work are:

- A novel method that, for the first time to our knowledge, combines privileged information and prototype learning to provide increased explanatory power for medical classification tasks.
- A prototype network architecture based on a capsule network that leverages the benefits of both techniques.
- An explainable solution outperforming state-of-the-art explainable and non-explainable methods on the LIDC-IDRI dataset.

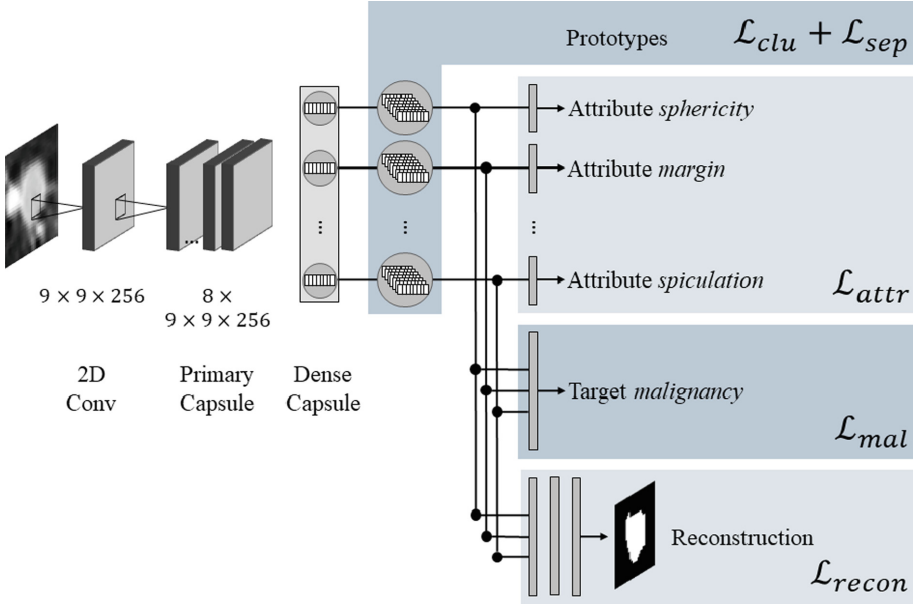
We provide the code with the model architecture and training algorithm of *Proto-Caps* on GitHub.

## 2 Methods

The idea behind our approach is to combine the potential of attribute and prototype learning for a powerful and interpretable learning system. For this, we use a capsule network of attribute capsules from which the target class is predicted. As the attribute prediction can also be susceptible to error, we use prototypes to explain the predictions made for each attribute. Based on [11], our approach, called *Proto-Caps*, consists of a backbone capsule network. The network is trained using multiple heads. An attribute head is used to ensure that each capsule represents a single attribute, a reconstruction head learns the original segmentation, and the main target prediction head learns the final classification. The model is extended by a prototype layer that provides explanations for each attribute decision. The overall architecture of *Proto-Caps* is shown in Fig. 1.

The **backbone** of our approach is a capsule network consisting of three layers: Features of the input image of size  $1 \times 32 \times 32$  are extracted by a 2D convolutional layer containing 256 kernels of size  $9 \times 9$ . We decided not to use 3D convolutional layers, as preliminary experiments showed only marginal differences (within std. dev. of results), but required significantly more computing time. The primary capsule layer then segregates low-level features into 8 different capsules, with each capsule applying 256 kernels of size  $9 \times 9$ . The final dense capsule layer consists of one capsule for each attribute and extracts high-level features, overall producing eight 16-dimensional vectors. These vectors form the starting point for the different prediction branches.

The **target head**, a fully connected layer, combines the capsule encodings. The loss function for the malignancy prediction was chosen according to LaLonde et al. [11], where the distribution of radiologist malignancy annotations is optimized with the Kullback-Leibler divergence  $\mathcal{L}_{mal}$  to reflect the inter-observer agreement and thus uncertainty. The **reconstruction branch** to predict the segmentation mask of the nodule consists of a simple decoder with three fully connected layers with the output filters 512, 1024, and the size of the resulting image  $1 \times 32 \times 32$ . The reconstruction loss  $\mathcal{L}_{recon}$  implements the mean square error between the output and the binary segmentation mask. It has been shown that incorporating reconstruction learning is beneficial to performance [11].



**Fig. 1. Proposed Model Architecture.** The backbone capsule network results in capsules representing predefined attributes. For each capsule, a set of prototypes is trained. To fit the attribute scores, the capsule vectors are fed through individual dense layers. The latent vectors of all capsules are being accumulated for a dense layer to predict a target score and for a decoder network to reconstruct the region of interest.

For the **attribute head**, we propose to use fully connected layers, instead of determining the attribute manifestation by the length of the capsule encoding, as was done previously [11]. Each capsule vector is processed by a separate linear layer to fit the respective attribute score. We formulate the attribute loss as

$$\mathcal{L}_{attr} = (1 - b) \sum_a \|Y_a - O_a\|^2, \quad (1)$$

where  $Y_a$  is the ground truth mean attribute score by the radiologists,  $O_a$  is the network score prediction for the  $a$ -th attribute, and  $b$  is a random binary mask allowing semi-supervised attribute learning.

Two **prototypes** are learned per possible attribute class, resulting in 8–12 prototypes per attribute (i.e. capsule). During the training, a combined loss function encourages a training sample to be close to a prototype of the correct attribute class and away from prototypes dedicated to others, similar to existing approaches [6]. Randomly initialized, the prototypes are a representative subset of the training dataset for each attribute after the training. For this, a cluster cost reduces the Euclidean distance of a sample’s capsule vector  $O_a$  to the nearest prototype vector  $p_j$  of group  $P_{a_s}$  which is dedicated to its correct attribute score.

$$\mathcal{L}_{clu} = \frac{1}{A} \sum_a \min_{p_j \in P_{a_s}} \|O_a - p_j\|_2. \quad (2)$$

In order to clearly distinguish between different attribute specifications, a separation loss is applied to increase the distance to the capsule prototypes that do not have the correct specification, limited by a maximum distance:

$$\mathcal{L}_{sep} = \frac{1}{A} \sum_a \min_{p_j \notin P_{a_s}} \max(0, \text{dist}_{max} - \|O_a - p_j\|_2). \quad (3)$$

Prototype optimization begins after 100 epochs. In addition to fitting the prototypes with the loss function, each prototype is replaced every 10 epochs by the most similar latent vector of a training sample. The original image of the training sample is stored and used for prototype visualization. During inference, the predicted attribute value is set to the ground truth attribute value of the closest prototype, ignoring the learned dense layers in the attribute head at this stage.

The overall loss function is the following weighted sum, where  $\lambda_{recon} = 0.512$  was chosen according to [11], and the prototype weights were chosen empirically:

$$\mathcal{L} = \mathcal{L}_{mal} + \lambda_{recon} \cdot \mathcal{L}_{recon} + \mathcal{L}_{attr} + 0.125 \cdot (\mathcal{L}_{clu} + 0.1 \cdot \mathcal{L}_{sep}) \quad (4)$$

### 3 Experiments

**Data.** The proposed approach is evaluated using the publicly available LIDC-IDRI dataset consisting of 1018 clinical thoracic CT scans from patients with Non-Small Cell Lung Cancer (NSCLC) [2,3]. Each lung nodule with a minimum size of 3 mm was segmented and annotated with a malignancy score ranging from 1-*highly unlikely* to 5-*highly suspicious* by one to four expert raters. Nodules were also scored according to their characteristics with respect to pre-defined attributes, namely subtlety (difficulty of detection, 1-*extremely subtle*, 5-*obvious*), internal structure (1-*soft tissue*, 4-*air*), pattern of calcification (1-*popcorn*, 6-*absent*), sphericity (1-*linear*, 5-*round*), margin (1-*poorly defined*, 5-*sharp*), lobulation (1-*no lobulation*, 5-*marked lobulation*), spiculation (1-*no spiculation*, 5-*marked spiculation*), and texture (1-*non-solid*, 5-*solid*). The `pylidc` framework [7] is used to access and process the data. The mean attribute annotation and the mean and standard deviation of the malignancy annotations are calculated. The latter was used to fit a Gaussian distribution, which serves as the ground truth label for optimization. Samples with a mean expert malignancy score of 3-*indeterminate* or annotations from fewer than three experts were excluded in consistency with the literature [8,9,11].

**Experiment Designs.** To ensure comparability with previous work [8,9,11], the main metric used is Within-1-Accuracy, where a prediction within one score is considered correct. Five-fold stratified cross-validation was performed using 10% of the training data for validation and the best run of three is reported.

The algorithm was implemented using the **PyTorch** framework version 1.13 and **CUDA** version 11.6. A learning rate of 0.5 was chosen for the prototype vectors and 0.02 for the other learnable parameters. The batch size was set to 128 and the optimizer was **ADAM** [10]. With a maximum of 1000 epochs, but stopping early if there was no improvement in target accuracy within 100 epochs, the experiments lasted an average of three hours on a GeForce RTX 3090 graphics card. The code is publicly available at <https://github.com/XRad-Ulm/Proto-Caps>.

Besides pure performance, the effect of reduced availability of attribute annotations was investigated. This was done by using attribute information only for a randomly selected fraction of the nodules during the training.

To investigate the effect of prototypes on the network performance, an ablation study was performed. Three networks were compared: **Proto-Caps** (proposed) including learning and applying prototypes during inference, **Proto-Caps<sub>w/o use</sub>** where prototypes are only learned but ignored for inference, and **Proto-Caps<sub>w/o learn</sub>** using the proposed architecture without any prototypes.

## 4 Results

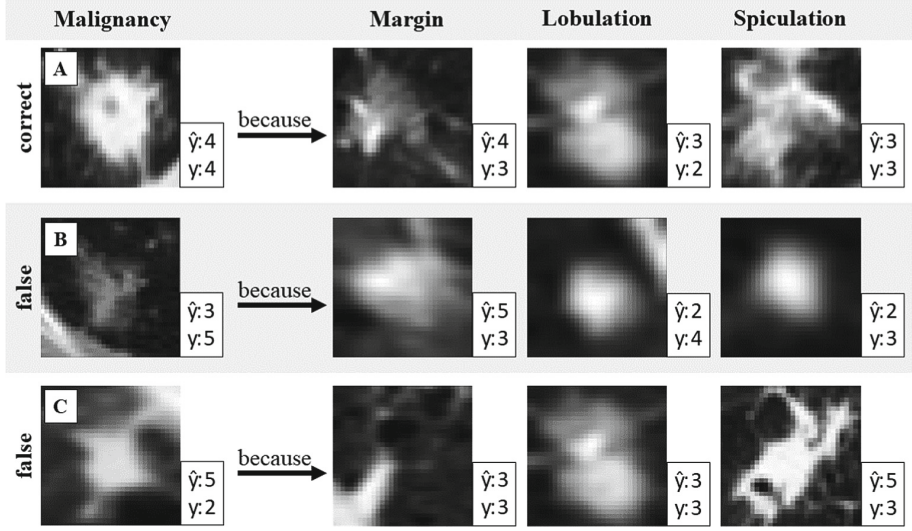
**Qualitative.** Figure 2 shows examples of model output. The predicted malignancy score is justified by the closest prototypical sample of a certain attribute. The respective original image for each attribute prototype is being saved during the training process and used for visualization during inference. In case B, there are large differences between the margin and lobulation prototype and the sample. Similarly, in case C, the spiculation prediction is very different from the sample.

During application, these discrepancies between the prototypes and the sample nodule raise suspicion, and help to assess the malignancy prediction. A quantitative evaluation of the relationship between correctness in attribute and in target prediction using logistic regression analysis shows a strong relationship between both with an accuracy of 0.93/0.1.

**Quantitative.** Table 1 shows the results of our experiments compared to other state-of-the-art approaches, with results taken from original reports. The accuracy of the proposed method exceeds previous work in both the malignancy and almost all attribute predictions, while modelling all given attributes.

Table 2 lists the results obtained when only fractions of the training samples come with attribute information. The experiments indicate that the performance of the given approach is maintained up to a fraction of 10%. Using no attribute annotations at all, i.e. no privileged information, achieves a similar performance, but results in a loss of explainability, as the high-level features extracted in the capsules are not understandable to humans. This result suggests that privileged

information here leads to an increase in interpretability for humans by providing attribute predictions and prototypes without interfering with the model performance.



**Fig. 2.** One correct and two wrongly predicted examples with exemplary attribute prototypes. Prediction  $\hat{y}$  and ground truth label  $y$  of malignancy and attribute respectively. Identifying false attribute predictions can help to identify misclassification in malignancy.

The ablation study shows no significant differences between the three models evaluated. For the malignancy accuracy, **Proto-Caps**<sub>w/o use</sub> and **Proto-Caps**<sub>w/o learn</sub> achieved  $\mu = 93.9\%$  ( $\sigma = 0.8$ ) and  $\mu = 93.7\%$  ( $\sigma = 1.1$ ), respectively. The average difference in attribute accuracy compared to the proposed methods is 1.7% and 1.5% better, respectively, and is more robust across experiments. The best result was obtained when the prototypes were learned but not used, possibly indicating that the prototypes may have a regularising effect during training, but further experiments are needed to confirm this due to the close results. To give an indication of the decoder performance, **Proto-Caps**<sub>w/o use</sub> achieved a dice score of 79.7%.

**Table 1.** Comparison with literature values of other works, attribute scores are reported if available. Mean  $\mu$  and standard deviation  $\sigma$  calculated from 5-fold experiments. Scores reported as Within-1-Accuracy, except for [16]. The best result is in bold.

	Attribute Prediction Accuracy in %								Malignancy
	Sub	IS	Cal	Sph	Mar	Lob	Spic	Tex	
<b>Non-explainable</b>									
3D-CNN+MTL [8]	–	–	–	–	–	–	–	–	90.0
TumorNet [9]	–	–	–	–	–	–	–	–	92.3
CapsNet [11]	–	–	–	–	–	–	–	–	77.0
<b>Explainable</b>									
HSCNN [16] (binary ACC.)	71.9	–	90.8	55.2	72.5	–	–	83.4	84.2
X-Caps [11]	<b>90.4</b>	–	–	85.4	84.1	70.7	75.2	93.1	86.4
Proto-Caps (proposed)	$\mu$ 89.1	<b>99.8</b>	<b>95.4</b>	<b>96.0</b>	<b>88.3</b>	<b>87.9</b>	<b>89.1</b>	<b>93.3</b>	<b>93.0</b>
	$\sigma$ 5.2	0.2	1.3	2.2	3.1	0.8	1.3	1.0	1.5

**Table 2.** Results of data reduction studies where attribute information was only available in fractions of the training dataset. Mean  $\mu$  and standard deviation  $\sigma$  calculated from 5-fold experiments. Scores reported as Within-1-Accuracy.

		Attribute Prediction Accuracy in %								Malignancy
		Sub	IS	Cal	Sph	Mar	Lob	Spic	Tex	
100 % attribute labels	$\mu$	89.1	99.8	95.4	96.0	88.3	87.9	89.1	93.3	93.0
	$\sigma$	5.2	0.2	1.3	2.2	3.1	0.8	1.3	1.0	1.5
10 % attribute labels	$\mu$	92.6	99.8	95.7	94.9	90.3	88.8	86.9	92.3	92.4
	$\sigma$	0.9	0.2	0.9	4.1	1.6	1.6	2.4	1.4	0.8
1 % attribute labels	$\mu$	91.0	99.8	92.8	95.5	79.9	85.7	85.6	91.2	90.2
	$\sigma$	4.5	0.2	1.4	2.3	13.1	4.4	6.8	1.7	1.1
0 % attribute labels	$\mu$	–	–	–	–	–	–	–	–	92.4
	$\sigma$	–	–	–	–	–	–	–	–	1.0

## 5 Discussion and Conclusion

We propose a new method, named *Proto-Caps*, which combines the advantages of privileged information, and prototype learning for an explainable network, achieving more than 6% better accuracy than the state-of-the-art explainable method. As shown by qualitative results (Fig. 2), the obtained prototypes can be used to detect potential false classifications. Our method is based on capsule networks, which allow prediction based on attribute-specific prototypes. Compared to class-specific prototypes, our approach is more specific and allows better interpretation of the predictions made. In summary, *Proto-Caps* outputs prediction results for the main classification task and for predefined attributes, and provides visual validation through the prototypical samples of the attributes.



The experiments demonstrate that it outperforms state-of-the-art methods that provide less explainability. Our data reduction studies show that the proposed solution is robust to the number of annotated examples, and good results are obtained even with a 90% reduction in privileged information. This opens the door for application to other datasets by reducing the additional annotation overhead. While we did see a reduction in performance with too few labels, our results suggest that this is mainly due to inhomogeneous coverage of individual attribute values. In this respect, it would be interesting to find out how a specific selection of the annotated samples, e.g. with extremes, affects the accuracies, especially since our results show that the overall performance is robust even when the attributes are not explicitly trained, i.e. without additional privileged information. Another area of research would be to explore other types of privileged information that require less extra annotation effort, such as medical reports, to train the attribute capsules. It would also be worth investigating more sophisticated 3D-based capsule networks.

In conclusion, we believe that the approach of leveraging privileged information with comprehensible architectures and prototype learning is promising for various high-risk application domains and offers many opportunities for further research.

**Acknowledgements.** This research was supported by the University of Ulm (Baustein, L.SBN.0214), and the German Federal Ministry of Education and Research (BMBF) within RACOON COMBINE “NUM 2.0” (FKZ: 01KX2121). We acknowledge the National Cancer Institute and the Foundation for the National Institutes of Health for the critical role in the creation of the publicly available LIDC/IDRI Dataset.

## References

1. Afshar, P., Mohammadi, A., Plataniotis, K.N.: Brain tumor type classification via capsule networks. In: 2018 25th IEEE International Conference on Image Processing (ICIP), pp. 3129–3133. IEEE (2018). <https://doi.org/10.1109/ICIP.2018.8451379>
2. Armato III, S.G., et al.: The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans. *Med. Phys.* **38**(2), 915–931 (2011). <https://doi.org/10.1118/1.3528204>, publisher: Wiley Online Library
3. Armato III, S.G., et al.: Data from LIDC-IDRI. The Cancer Imaging Archive (2015). <https://doi.org/10.7937/K9/TCIA.2015.LO9QL9SX>, [Data set]
4. Barnett, A.J., et al.: A case-based interpretable deep learning model for classification of mass lesions in digital mammography. *Nat. Mach. Intell.* **3**(12), 1061–1070 (2021). <https://doi.org/10.1038/s42256-021-00423-x>
5. Bien, J., Tibshirani, R.: Prototype selection for interpretable classification. *Ann. Appl. Stat.* **5**(4), 2403–2424 (2011). <https://doi.org/10.1214/11-AOAS495>
6. Chen, C., Li, O., Tao, D., Barnett, A., Rudin, C., Su, J.K.: This looks like that: deep learning for interpretable image recognition. In: *Advances in Neural Information Processing Systems 32* (2019)

7. Hancock, M.C., Magnan, J.F.: Lung nodule malignancy classification using only radiologist-quantified image features as inputs to statistical learning algorithms: probing the lung image database consortium dataset with two statistical learning methods. *J. Med. Imaging*. **3**(4), 044504–044504 (2016). <https://doi.org/10.1117/1.JMI.3.4.044504>
8. Hussein, S., Cao, K., Song, Q., Bagci, U.: Risk stratification of lung nodules using 3D CNN-based multi-task learning. In: Niethammer, M., et al. (eds.) *IPMI 2017*. LNCS, vol. 10265, pp. 249–260. Springer, Cham (2017). [https://doi.org/10.1007/978-3-319-59050-9\\_20](https://doi.org/10.1007/978-3-319-59050-9_20)
9. Hussein, S., Gillies, R., Cao, K., Song, Q., Bagci, U.: TumorNet: Lung nodule characterization using multi-view convolutional neural network with Gaussian process. In: *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*, pp. 1007–1010. IEEE (2017). <https://doi.org/10.1109/ISBI.2017.7950686>
10. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014)
11. LaLonde, R., Torigian, D., Bagci, U.: Encoding visual attributes in capsules for explainable medical diagnoses. In: Martel, A.L., et al. (eds.) *MICCAI 2020*. LNCS, vol. 12261, pp. 294–304. Springer, Cham (2020). [https://doi.org/10.1007/978-3-030-59710-8\\_29](https://doi.org/10.1007/978-3-030-59710-8_29)
12. Li, O., Liu, H., Chen, C., Rudin, C.: Deep learning for case-based reasoning through prototypes: a neural network that explains its predictions. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, issue 1 (2018)
13. Pan, Y., Yao, T., Li, Y., Wang, Y., Ngo, C.W., Mei, T.: Transferrable prototypical networks for unsupervised domain adaptation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2239–2247 (2019). <https://doi.org/10.1109/CVPR.2019.00234>
14. Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* **1**(5), 206–215 (2019). <https://doi.org/10.1038/s42256-019-0048-x>
15. Sabour, S., Frosst, N., Hinton, G.E.: Dynamic routing between capsules. In: *Advances in Neural Information Processing Systems 30* (2017)
16. Shen, S., Han, S.X., Aberle, D.R., Bui, A.A., Hsu, W.: An interpretable deep hierarchical semantic convolutional neural network for lung nodule malignancy classification. *Expert Syst. Appl.* **128**, 84–95 (2019). <https://doi.org/10.1016/j.eswa.2019.01.048>, publisher: Elsevier
17. Snell, J., Swersky, K., Zemel, R.: Prototypical networks for few-shot learning. In: *Advances in Neural Information Processing Systems 30* (2017)
18. Sun, S., Sun, Q., Zhou, K., Lv, T.: Hierarchical attention prototypical networks for few-shot text classification. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 476–485 (2019). <https://doi.org/10.18653/v1/D19-1045>
19. Vapnik, V., Izmailov, R., et al.: Learning using privileged information: similarity control and knowledge transfer. *J. Mach. Learn. Res.* **16**(1), 2023–2049 (2015)
20. Vapnik, V., Vashist, A.: A new learning paradigm: learning using privileged information. *Neural Netw.* **22**(5), 544–557 (2009). <https://doi.org/10.1016/j.neunet.2009.06.042>, advances in Neural Networks Research: IJCNN2009
21. Wang, M., Guo, Z., Li, H.: A dynamic routing CapsNet based on increment prototype clustering for overcoming catastrophic forgetting. *IET Comput. Vis.* **16**(1), 83–97 (2022). <https://doi.org/10.1049/cvi2.12068>, publisher: Wiley Online Library

22. Xu, W., Xian, Y., Wang, J., Schiele, B., Akata, Z.: Attribute prototype network for zero-shot learning. *Adv. Neural. Inf. Process. Syst.* **33**, 21969–21980 (2020)
23. Zheng, H., Fu, J., Mei, T., Luo, J.: Learning multi-attention convolutional neural network for fine-grained image recognition. In: *Proceedings of the IEEE International Conference on Computer Vision* pp. 5209–5217 (2017). <https://doi.org/10.1109/ICCV.2017.557>
24. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2921–2929 (2016). <https://doi.org/10.1109/CVPR.2016.319>