Enhancing Prototypical Space for Interpretable Image Classification

Jae Soon Baik*, Kyoung Ok Yang*, and Jun Won Choi†
Hanyang University
Seoul, South Korea
Email: {jsbaik, koyang}@spa.hanyang.ac.kr, junwchoi@hanyang.ac.kr

***Abstract*—In the Industry 4.0 era, machine learning and artificial intelligence have emerged as influential drivers in the domain of information and communication technology (ICT). This fusion of human interaction and technological progress emphasizes the necessity for artificial intelligence systems that provide transparent outcomes. The introduction of explainable artificial intelligence (XAI) plays a crucial role in establishing trust and facilitating the broader integration of AI. However, the earlier prototype-based XAI models encounter a constraint in acquiring a meaningful representation for all patches in the prototypical space. This limitation arises from its inherent update process, wherein the update gradient is only transmitted through the most active patch of the input image. To address this limitation, we introduce an innovative XAI model that introduces a novel reconstruction loss implemented by adopting a Variational autoencoder (VAE). This reconstruction loss offers insights into the rationale underlying the prototypical space and similarity mechanism within part prototypical-based XAI model, thereby enhancing comprehension of its fundamental principles. Extensive experimental results demonstrate the effectiveness of our method over existing approaches through evaluation on the widely used CUB-200-2011 dataset.**

***Index Terms*—Explainable AI, Part Prototype-based Model, Image Processing, Reconstruction Loss**

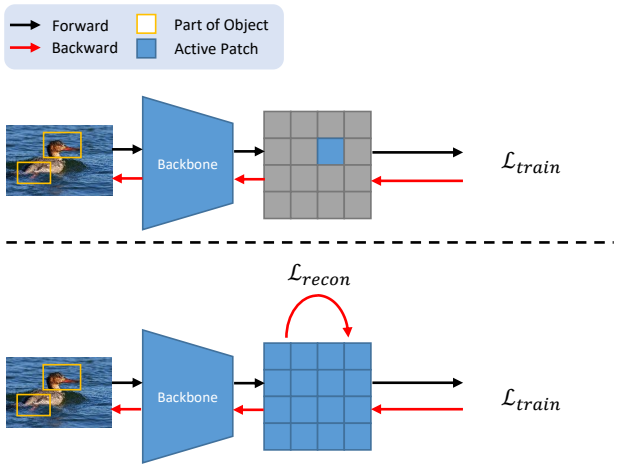

Fig. 1. **Conceptual illustration of (top) conventional part prototype-based method and (bottom) the proposed method.** In the conventional part prototype-based method, gradients are passed through only the most active patches to update the model. Our method, on the other hand, employs the proposed reconstruction loss to enable more effective gradient updating.

I. Introduction

In the ongoing technological revolution of Industry 4.0, machine learning (ML) and artificial intelligence (AI) have emerged as the driving forces shaping every aspect of information and communication technology (ICT) [1]. These technologies cover a wide range of tasks from intricate data processing [2] to advanced methodologies like machine learning operations (MLOps) [3]. Amidst these developments, the interaction between human and technological systems is becoming increasingly important in the ICT domains, particularly in critical domains such as healthcare [4], autonomous driving [5], and robotics [6], where high levels of trust are imperative. Consequently, there has been a noticeable increase in the demand for AI systems capable of generating results that are both comprehensible and explainable for humans.

In the field of image recognition tasks, a pioneering study has introduced an innovative prototype-based *explainable artificial intelligence* (XAI) model referred to as the *"prototypical part network"* (ProtoPNet) [7]. ProtoPNet discriminates novel objects by assessing the relevance of input images to learned part prototypes that represent specific object parts in the input image. This process enhances human comprehension by providing transparency into the degree of involvement of different object parts when classifying novel instances.

However, ProtoPNet requires M part prototypes for each class, which has the disadvantage that the number of prototypes increases with the number of classes. To overcome this disadvantage, ProtoTree [8] and ProtoPool [9] proposed a class-agnostic method. ProtoTree reduced the number of prototypes by using a decision tree explanation technique. ProtoPool departed from the conventional practice of hard assignment of prototypes to classes. Instead, it employed a soft assignment mechanism that selects a set of related prototypes from the pool of class-agnostic prototypes. This approach allowed part prototypes to be shared for each new instance, reducing the number of prototypes involved in the reasoning process.

In part prototype-based XAI research, the additional interpretable process leads to a decline in performance due to their distinctive inference procedure. In this regard, as observed in approaches such as TesNet [10] and ProtoPool [9], multiple techniques have been proposed to enhance the prototypical

*Equal contribution
†Corresponding author

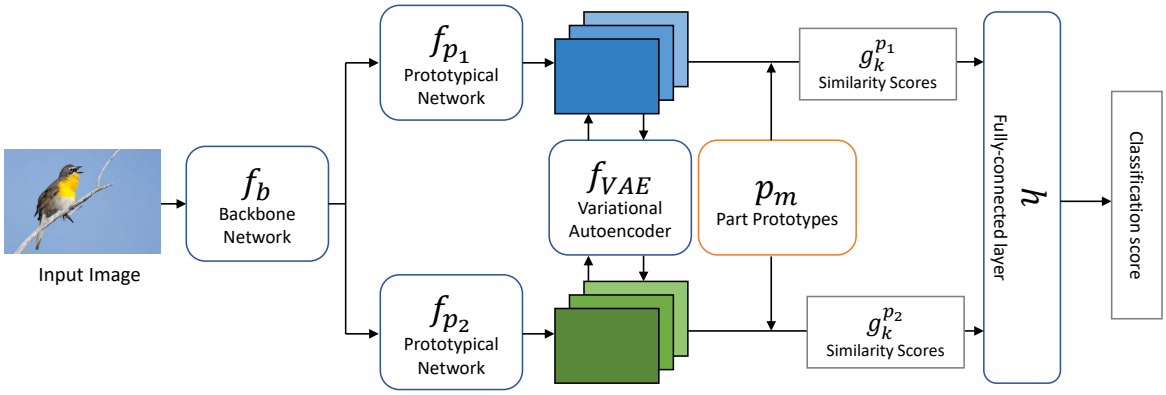

Fig. 2. **Overview of the proposed method**. The proposed method consists of a convolutional backbone network f_b, multiple prototypical layer f_{p_i}, variational autoencoder f_{VAE}, and fully connected layer h.

space, the embedding space where comparisons with part prototypes are conducted. TesNet boosted the transparency of the embedding space by introducing the Grassmann manifold. In this framework, the basis concepts of a specific category are mutually perpendicular, increasing the clarity of the basis concepts and, in turn, enhancing the overall performance. ProtoPool utilized focal similarity to increase the distance between active and inactive patches, resulting in the XAI model focusing on salient features rather than the background. However, since focal similarity primarily aims to enhance the gap between patches in the prototypical space, it does not guarantee that all patches sufficiently learn meaningful semantic information. As a consequence, ProtoPool faced challenges in consistently conducting reliable comparisons between part prototypes and the embedded features represented in the prototype space.

To address this issue, we present a novel framework that improves the prototypical space. The conceptual illustration of the proposed framework is demonstrated in Fig. 1. As shown in Fig. 1 (top), conventional part prototype-based methods propagate gradients only through the most active patches in the image. Thus, the remaining inactive patches are limited to acquiring appropriate semantic features. In contrast, our approach introduces a novel structure with reconstruction loss that enhances the semantic representation of inactive features in the prototypical space, as shown in Fig. 1 (bottom). The reconstruction loss allows inactive patches to reconstruct embedding vectors in a different prototypical space so that all patches, even those that are inactive, can learn significant semantic information about the object that was mostly learned and represented by the most active patches. The reconstruction loss improves the comparison between the embedding vector and the part prototype in the prototypical space, leading to more accurate results.

To implement the aforementioned reconstruction loss, we employ two modules, as illustrated in Fig. 2. Firstly, we incorporate two prototypical layers, each using different parameters, to extract different embedding spaces using the multi-view approach. Each embedding space mutually reconstructs the other, facilitating the acquisition of meaningful semantic features for all patches. We also employ a Variational Autoencoder (VAE) to calculate the reconstruction loss. The VAE takes embedding features from one embedding space as input and uses them to reconstruct the embedding features in the other embedding space. In conclusion, we can effectively improve both the classification performance and the transparency of the inference process by accurately obtaining semantic information about parts of the object that were inadequately trained during the training phase.

We evaluated the performance using the commonly used CUB-200-2011 dataset in the field of XAI. Our approach showed a significant improvement, achieving a performance gain of 1.7% over the ProtoPool baseline method [9]. The proposed method outperformed the performance of other XAI methods by a substantial margin. These experimental results demonstrate the efficiency of the proposed method to improve the prototypical space.

II. Related Work

Prototype-based XAI is an approach that seeks to enhance the transparency and interpretability of deep neural networks by utilizing prototypes to explain their decision-making processes. In this respect, it's noteworthy that prototype-based XAI primarily aligns with the category of example-based XAI methodologies. This alignment is attributed to the core mechanism of prototype-based XAI, which involves tackling an optimization challenge to identify points within the dataset that possess minimal distance from all other data points. These techniques are discerned through the resolution of an optimization problem aimed at determining the point with the minimum distance to all other points within the dataset [11]–[13]. The deployment of prototypes can notably enhance the interpretability of a classification model by effectively

illustrating the distinctions inherent in the representative data points for each individual class.

However, prototypes might be insufficient due to their limited capacity to provide explanations in localized contexts. ProtoPNet [7] addresses this by incorporating self-explanation mechanisms within specific regions through the utilization of part prototypes. ProtoPNet has played a pivotal role in inspiring the emergence of prototype-based XAI models. For instance, TesNet [10] constructs the latent space on a Grassmann manifold without resorting to prototype reduction. Similarly, models such as ProtoPShare [14] and ProtoTree [8] have been proposed to optimize the number of prototypes employed in the classification process. ProtoPool [9] implemented a soft assignment strategy, characterized by the distribution spanning the prototype set, which yields a reduction in the prototype count, leading to enhanced interpretability and streamlined training. In this paper, our goal is to enhance the effectiveness of the primary prototypes in representing their respective concepts. We achieve this by allowing the learning process to consider not only the inherent characteristic of the prototype itself but also the relevant surrounding information that is crucial for its creation.

III. Proposed Method

A. Overview of the Proposed Method

The overview of the proposed method is illustrated in Figure 2. In this paper, we implement our method using ProtoPool as a baseline interpretable method. Our proposed method consists of several components: a CNN backbone network f_b, two prototypical layers f_{p_i} indexed by i, a Variational Autoencoder (VAE) f_{VAE} for reconstruction loss, and a fully connected layer h as a classifier.

Given an input image $I \in \mathbb{R}^{H_0 \times W_0 \times 3}$, the CNN backbone network produces a feature map $Z = f_b(x) \in \mathbb{R}^{H \times W \times D}$. Similar to ProtoPool, we can consider the backbone feature as the set of patches $Z = \{z_t \in f_b(x); z_t \in \mathbb{R}^D, t = 1, ..., H \cdot W\}$. Two distinct prototypical layers have K slots for each layer and share the learnable part prototypes $\mathcal{P} = \{p_m \in \mathbb{R}^{H_p \times W_p \times D}\}_{m=1}^{M}$, where the dimensions of H_p and W_p are both set to 1. Then, the prototypical layers f_{p_i} aggregate the similarity scores as $g_k^{p_i} = \sum_{m=1}^{M} q_k g_m^{p_i}$, where $q_k \in \mathbb{R}^M$ denotes the distribution obtained by each slot and $g_m^{p_i}$ represents a similarity score obtained in the i-th prototypical layer compared with the mth part prototype. Similar to previous works [7], the similarity score with respect to input patch z can be calculated as

$$g_m^{p_i}(z) = log\left(\frac{(||f_{p_i}(z) - p_m||_2^2 + 1)}{(||f_{p_i}(z) - p_m||_2^2 + \epsilon)}\right), \tag{1}$$

where $f_{p_i}(z)$ represents the prototypical feature vector, p_m denotes the part prototype and ϵ denotes the constant value for avoiding numerical instability. The final similarity score is determined as $g_m^{p_i} = \max_{z \in Z} g_m^{p_i}(z)$. A high activation value of $g_m^{p_i}$ indicates that the active patch and prototype p_m are similar. Lastly, we obtain the K similarity scores $g_k^{p_i}$ from each of the prototypical layers f_{p_i} and feed them into the classifier h to compute the classification score.

B. Reconstruction Loss

In this study, we propose a novel VAE-driven reconstruction loss for enhancing the prototypical space. The proposed VAE is fed the feature vector from one prototypical space and then reconstructs the feature in another prototypical space. The reconstruction loss $\mathcal{L}_{recon}$ minimizes the variational lower bound on the marginal likelihood can be described as

$$\begin{aligned}\mathcal{L}_{recon}^1 = \mathbb{E}[log_{p_\theta}(f_{p_2}(z)|z_{vae})] \\ - \beta \mathcal{D}_{KL}(q_\phi(z_{vae}|f_{p_1}(z))||p(z_{vae}))\end{aligned} \tag{2}$$

$$\begin{aligned}\mathcal{L}_{recon}^2 = \mathbb{E}[log_{p_\theta}(f_{p_1}(z)|z_{vae})] \\ - \beta \mathcal{D}_{KL}(q_\phi(z_{vae}|f_{p_2}(z))||p(z_{vae}))\end{aligned} \tag{3}$$

where q_ϕ and p_θ are the encoder and decoder of VAE parameterized by ϕ and θ, respectively. $p(z_{vae})$ denotes the Gaussian prior distribution for VAE, β denotes the Lagrangian parameter for optimization, and $\mathcal{D}_{KL}$ denotes the Kullback–Leibler (KL) divergence.

Total Loss Then, the final loss is obtained as

$$\begin{aligned}\mathcal{L} = \mathcal{L}_{entropy} + \lambda_{clst}\mathcal{L}_{clst} + \lambda_{sep}\mathcal{L}_{sep} \\ + \lambda_{orth}\mathcal{L}_{orth} + \lambda_{l_1}\mathcal{L}_{l_1} \\ + \lambda_{recon}\mathcal{L}_{recon}^1 + \lambda_{recon}\mathcal{L}_{recon}^2\end{aligned} \tag{4}$$

where $\mathcal{L}_{entropy}$ denotes the cross-entropy loss, $\mathcal{L}_{clst}$ denotes the clustering loss, $\mathcal{L}_{sep}$ denotes the separation loss, $\mathcal{L}_{orth}$ denotes the orthogonality loss, and $\mathcal{L}_{l_1}$ denotes the l_1 regularization applied to the classifier. Except for the newly proposed reconstruction loss, the loss functions and their coefficients are the same as those used in the baseline method [9].

C. Prototype Projection

After the training process, the prototypes are mapped onto the input patch for visualization and transparency. In this projection step, the prototypes are projected to the nearest active input patch as

$$p_m \leftarrow \arg\min_{z \in Z} ||f_{p_i}(z) - p_m||_2. \tag{5}$$

Through this process, we gain insight into the parts of the object that contributed to the classification of the input image. We follow prototype visualization methodology as used in [7], [9].

IV. Experiments

A. Experimental Setups

We evaluated the model on CUB-200-2011 dataset, which contains 5,994 training images and 5,794 test images. CUB-200-2011 dataset provides 200 bird species. Following the experimental setup in ProtoPNet [7], The model was trained on cropped images obtained from bounding box annotations within the dataset. Subsequently, we used offline augmentation, including shearing, image rotation, skewing, and horizontal flipping, during the training process. For CNN backbone

TABLE I
COMPARISON OF THE PROPOSED METHOD ON CUB-200-2011 DATASET. INR50 REPRESENTS THE RESNET-50 [16] PRETRAINED ON THE INATURALIST DATASET [15].

Method	Arch.	# of Proto.	Acc (%)
ProtoTree	iNR50	202	82.2±0.7
ProtoPool		202	85.5±0.1
Ours		202	**87.2±0.3**

network, we used ResNet-50 network initialized with the weight pretrained on iNaturalist 2018 dataset [15]. Following the configuration used in [9], the coefficients for each loss function were set to as follows: $\lambda_{entropy} = 1.0, \lambda_{clst} = 0.8, \lambda_{sep} = -0.08, \lambda_{orth} = 1.0$, and $\lambda_{l_1} = 0.0001$. For the reconstruction loss, we used a coefficient λ_{recon} of 0.2. The Lagrangian parameter β was set to 1.

B. Experimental Results

Table I presents the performance of the proposed method compared to other prototype-based XAI models. The proposed method achieves 5.0% and 1.7% better performance over ProtoTree and ProtoPool, respectively. These results demonstrate the efficacy of the proposed method, which enhances the prototypical space by adopting the reconstruction loss. This approach allows meaningful semantic information to be learned from inactive patches, effectively improving the transparency of the part prototype-based XAI model.

V. CONCLUSION

In this study, we introduced a novel XAI model with VAE-driven reconstruction loss to enhance the prototypical space. While conventional part prototype-based methods learn semantic information from the only active patch due to their distinctive inference procedure, applying the proposed reconstruction loss allows all patches in the prototypical space to learn meaningful semantic information, enhancing the transparency and the final performance. The experiments conducted on CUB-200-2011 demonstrated that the proposed method achieved significant performance improvements over the existing part prototype-based XAI models. In future work, we plan to incorporate the insight of concept-based learning to further improve the prototypical space and interpretability of the XAI model.

VI. ACKNOWLEDGEMENTS

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.2020-0-01373, Artificial Intelligence Graduate School Program (Hanyang University))

REFERENCES

[1] S. I. Tay, T. Lee, N. Hamid, and A. N. A. Ahmad, "An overview of industry 4.0: Definition, components, and government initiatives," *Journal of Advanced Research in Dynamical and Control Systems*, vol. 10, no. 14, pp. 1379–1387, 2018.

[2] Y. Zhong, L. Chen, C. Dan, and A. Rezaeipanah, "A systematic survey of data mining and big data analysis in internet of things," *The Journal of Supercomputing*, vol. 78, no. 17, pp. 18 405–18 453, 2022.

[3] D. Kreuzberger, N. Kühl, and S. Hirschl, "Machine learning operations (mlops): Overview, definition, and architecture," *IEEE Access*, 2023.

[4] M. A. Ahmad, C. Eckert, and A. Teredesai, "Interpretable machine learning in healthcare," in *Proceedings of the 2018 ACM international conference on bioinformatics, computational biology, and health informatics*, 2018, pp. 559–560.

[5] S. Atakishiyev, M. Salameh, H. Yao, and R. Goebel, "Explainable artificial intelligence for autonomous driving: A comprehensive overview and field guide for future research directions," *arXiv preprint arXiv:2112.11561*, 2021.

[6] E. Puiutta and E. M. Veith, "Explainable reinforcement learning: A survey," in *International cross-domain conference for machine learning and knowledge extraction*. Springer, 2020, pp. 77–95.

[7] C. Chen, O. Li, D. Tao, A. Barnett, C. Rudin, and J. K. Su, "This looks like that: deep learning for interpretable image recognition," *Advances in neural information processing systems*, vol. 32, 2019.

[8] M. Nauta, R. Van Bree, and C. Seifert, "Neural prototype trees for interpretable fine-grained image recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 14 933–14 943.

[9] D. Rymarczyk, Ł. Struski, M. Górszczak, K. Lewandowska, J. Tabor, and B. Zieliński, "Interpretable image classification with differentiable prototypes assignment," in *European Conference on Computer Vision*. Springer, 2022, pp. 351–368.

[10] J. Wang, H. Liu, X. Wang, and L. Jing, "Interpretable image recognition by constructing transparent embedding space," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 895–904.

[11] Y. Ming, P. Xu, F. Cheng, H. Qu, and L. Ren, "Protosteer: Steering deep sequence model with prototypes," *IEEE transactions on visualization and computer graphics*, vol. 26, no. 1, pp. 238–248, 2019.

[12] K. S. Gurumoorthy, A. Dhurandhar, G. Cecchi, and C. Aggarwal, "Efficient data representation by selecting prototypes with importance weights," in *2019 IEEE International Conference on Data Mining (ICDM)*. IEEE, 2019, pp. 260–269.

[13] O. Li, H. Liu, C. Chen, and C. Rudin, "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, no. 1, 2018.

[14] D. Rymarczyk, Ł. Struski, J. Tabor, and B. Zieliński, "Protopshare: Prototypical parts sharing for similarity discovery in interpretable image classification," in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 2021, pp. 1420–1430.

[15] G. Van Horn, O. Mac Aodha, Y. Song, Y. Cui, C. Sun, A. Shepard, H. Adam, P. Perona, and S. Belongie, "The inaturalist species classification and detection dataset," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8769–8778.

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.