Language Model Meets Prototypes: Towards Interpretable Text Classification Models through Prototypical Networks

Ximing Wen

Drexel University, Philadephia, USA
xw384@drexel.edu

Abstract

Pretrained transformer-based Language Models (LMs) are well-known for their ability to achieve significant improvement on NLP tasks, but their *black-box* nature, which leads to a lack of interpretability, has been a major concern. My dissertation focuses on developing intrinsically interpretable models when using LMs as encoders while maintaining their superior performance via prototypical networks. I initiated my research by investigating enhancements in performance for interpretable models of sarcasm detection. My proposed approach focuses on capturing sentiment incongruity to enhance accuracy while offering instance-based explanations for the classification decisions. Later, I develop a novel *white-box* multi-head graph attention-based prototype network designed to explain the decisions of text classification models without sacrificing the accuracy of the original black-box LMs. In addition, I am working on extending the attention-based prototype network with contrastive learning to redesign an interpretable graph neural network, aiming to enhance both the interpretability and performance of the model in document classification.

Background

Deep learning models, especially transformer-based Language Models (LMs) have significantly contributed to advancements in Natural Language Processing (NLP), offering encoders as powerful tools for text classification. However, despite their *state-of-art* performance, their complexity and *black-box* nature obscure the decision-making process and hinder their interpretability. Prototype networks, serving as a *white-box* framework where decisions are derived from similarity scores to instance-level prototypes, were initially proposed as an interpretable architecture in the image domain (Li et al. 2018; Chen et al. 2019). This approach was later adapted to the NLP domain (Ming et al. 2019; Hong, Wang, and Baek 2023).

Based on the classic framework of prototype learning (Datta and Kibler 1995), the prototype approach learns prototype vectors through training, projected onto representative cases from previous observations, to explain decisions more intuitively as shown in Figure 1. However, existing approaches for text classification still have performance gaps compared to the original *black-box* model. Moreover, in the context of document classification utilizing Graph Neural Networks (Gori, Monfardini, and Scarselli 2005), enhancing the graph prototypes to yield accurate and trustworthy explanations remains a significant challenge.

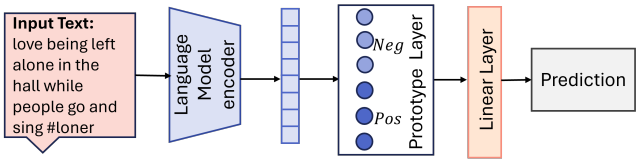

Figure 1: Illustration of prototype architecture for text classification

Goal of the Dissertation

My dissertation aims to minimize the performance gap of intrinsically interpretable models using prototype network networks in conjunction with LM encoders. I proposed a graph-attention-based prototype network framework to better learn the relatedness (i.e., similar semantic meaning) between the input and prototypes. This framework is applied and experimented with in the context of both sentence-based text classification and graph-based document classification tasks.

Contributions

A Transformer and Prototype Interpretable Model for Contextual Sarcasm Detection

My research began with an exploration aimed at minimizing the performance gap in interpretable prototype text classification, focusing initially on the task of sarcasm detection. I proposed an interpretable multi-view framework that integrates semantic and sentiment embeddings from pretrained LM encoders. Existing explainable methods, including attention-based and post-hoc approaches, generate word-level explanations, which may attribute similar importance to various words in contexts devoid of strong sentiment cues, particularly when sarcasm is conveyed through analogy rather than direct expression. The approach diverges from these methods by leveraging semantic prototypes to provide sentence-level, human-readable explanations. Furthermore, I innovate with the design of an incongruity loss,

which employs sentiment prototypes to discern implicit and explicit sentiments within textual expressions, thereby enhancing predictive accuracy. This methodology achieves *state-of-the-art* performance on benchmark datasets.

Status: Completed. Submitted to TACL.

Graph-attention Enhanced Prototype Network for Text Classification

I develop a novel white-box Multi-head Graph Attention-based prototype framework designed to explain the decisions of text classification models built with LM encoders. (Wen, Tan, and Weber 2024).The approach incorporates a prototype layer on top of a fine-tuned LM and utilizes multi-head graph attention (Velickovic et al. 2017) to efficiently learn relatedness by selectively constructing edges between encoded representations and their neighboring prototypes. In the reference time, the decision is solely based on the edge weights computed by each attention head. Different from other prototype networks that use heuristic metrics such as cosine similarity to learn relateness, I leverage graph attention network (GAT), which is known for its ability to capture the importance of neighboring nodes in a graph, enabling more effective learning for each prototype.

Performance: Extensive comparison experiments are conducted with variations of prototype networks on five public benchmark datasets including binary, four-label, and ten-label classification. I also experimented with multiple LMs as encoders. The approach achieved the best preformance compared with all prototype networks on all datasets. Compared with the original *black-box* models, the proposed approach either achieves the best performance, or the performance gap is within 0.3%.

Interpretibility: The case study shows different attention heads could capture different semantic aspects in the input sentence and activate the corresponding prototypes. Moreover, the trained prototype vectors are visualized within the training data space with t-SNE. It is found that the prototype vectors are evenly distributed, indicating that the space formed by these vectors can span over the space so a limited number of prototypes can be used to represent any datapoint. The percentage of distinguished prototypes is between 90%-95% when the number of prototypes varies from 10 to 40, suggesting the framework is robust in generating representative prototypes with respect to the change of hyperparameters.

Status: Completed. Accepted by COLING 2025.

Attention Enhanced Prototype Graph Neural Networks with Contrastive Learning

My next research plan aims to extend the attention-based prototype network delineated in (Wen, Tan, and Weber 2024) to the domain of document classification leveraging graph neural networks (GNNs). Contrary to sentence-based text classification, graph-based document classification conceptualizes documents as nodes within a graph, encapsulating relationships such as citations, co-authorships, and keyword associations as edges. The methodology employs transformer-based language models to initiate document node embeddings, while GNNs are utilized to derive comprehensive graph embeddings. Prototype vectors, initially randomly initialized, represent the graph embeddings associated with various classes. Previous endeavors, such as ProtoGNN (Zhang et al. 2022), utilize heuristic distances to measure the similarity between prototypes and input vectors; I hypothesize that the attention-based prototype network harbors the potential to enhance performance in this context. Moreover, existing approaches use proximity loss to ensure the prototypes can represent real training cases. However, this is more challenging for graph prototypes due to the complexity of graph structure. To address this, I propose the integration of contrastive learning, specifically GraphCL (You et al. 2020), to compel prototypes to distill and embody salient graph patterns, thereby augmenting interpretability.

Status: In Progress. Planned to be finished by 04/2024

References

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

Datta, P.; and Kibler, D. 1995. Learning prototypical concept descriptions. In *Machine Learning Proceedings 1995*, 158–166. Elsevier.

Gori, M.; Monfardini, G.; and Scarselli, F. 2005. A new model for learning in graph domains. In *Proceedings. 2005 IEEE international joint conference on neural networks, 2005.*, volume 2, 729–734. IEEE.

Hong, D.; Wang, T.; and Baek, S. 2023. ProtoryNet-interpretable text classification via prototype trajectories. *Journal of Machine Learning Research*, 24(264): 1–39.

Li, O.; Liu, H.; Chen, C.; and Rudin, C. 2018. Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Ming, Y.; Xu, P.; Qu, H.; and Ren, L. 2019. Interpretable and steerable sequence learning via prototypes. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 903–913.

Velickovic, P.; Cucurull, G.; Casanova, A.; Romero, A.; Lio, P.; Bengio, Y.; et al. 2017. Graph attention networks. *stat*, 1050(20): 10–48550.

Wen, X.; Tan, W.; and Weber, R. O. 2024. GAProtoNet: A Multi-head Graph Attention-based Prototypical Network for Interpretable Text Classification. *arXiv preprint arXiv:2409.13312*.

You, Y.; Chen, T.; Sui, Y.; Chen, T.; Wang, Z.; and Shen, Y. 2020. Graph contrastive learning with augmentations. *Advances in neural information processing systems*, 33: 5812–5823.

Zhang, Z.; Liu, Q.; Wang, H.; Lu, C.; and Lee, C. 2022. Protgnn: Towards self-explaining graph neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 9127–9135.