Explanation-Guided Training for Cross-Domain Few-Shot Classification

Jiamei Sun[1], Sebastian Lapuschkin[2], Wojciech Samek[2], Yunqing Zhao[1], Ngai-Man Cheung[1], Alexander Binder[1]
[1]Information System of Technology and Design, Singapore University of Technology and Design
[2]Department of Video Coding & Analytics, Fraunhofer Heinrich Hertz Institute, Berlin, Germany

Abstract

- In cross-domain few-shot classification (CD-FSC), we need to address not only the issue of **limited labeled data** in each class but also the **domain shift** between training and test domains.
- The use cases of **explanations** are still worthy to explore.
- We consider the question of whether **explanations** are suitable to **improve model performance** in small sample size regimes such as **few-shot classification**.

Introduction

The domain shift problem in FSC: FSC models are commonly evaluated using a test dataset originating from the **same** domain as the training dataset. These methods will meet difficulties in cases with the **domain shift between the training data (source domain) and the test data (target domain)**.

Explanation properties: To our best knowledge, there have not been explanation methods specially designed for FSC models. Favorable properties: **per-sample basis, low cost, no additional layers or trainable parameters, scores are related to the importance of a neuron.**

The contributions of this paper:

a) We derive explanations for FSC models using LRP.
b) We investigate the potential of improving model performance using explanations in the training phase under few-shot settings.
c) We propose an explanation-guided training strategy to tackle the domain shift problem in FSC.

Explanation-Guided Training

Key idea:

Re-weighting intermediate features using explanation scores.

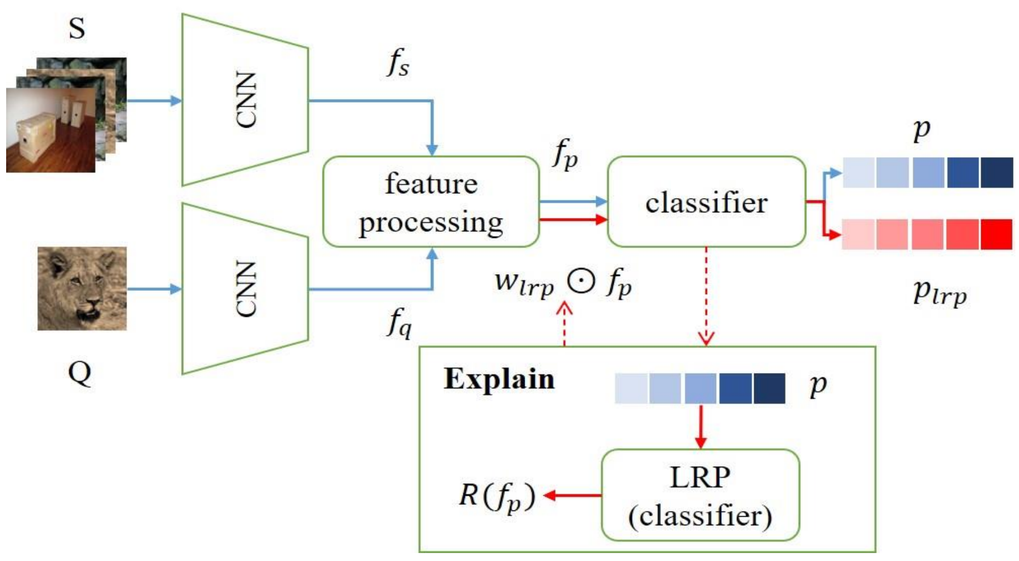

Figure 1. Explanation-guided training

Step1: One forward-pass through the model and obtain the prediction p.

Step2: Explaining the classifier.

- Initialize the relevance scores of the target labels.
  logits → neural networks classifiers
  logit function → non-parametric classifiers

$$P(y_c|f_p) = \frac{\exp(\beta \cdot cs_c(f_p))}{\sum_{k=1}^{K} \exp(\beta \cdot cs_k(f_p))}$$

$$R_c = log\left(\frac{P(y_c|f_p)}{1 - P(y_c|f_p)}(K-1)\right)$$

- Apply LRP on the classifier to obtain the relevance score of the intermediate feature $R(f_p)$, relying on ϵ-rule and $\alpha\beta$-rule[1].

Step3: Compute LRP-weighted features.

$$w_{lrp} = 1 + R(f_p)$$
$$f_{p-lrp} = w_{lrp} \odot f_p$$

Step4: Training with LRP-weighted features.

$$L = \xi L_{ce}(y, p) + \lambda L_{ce}(y, p_{lrp})$$

Experiment

- The performance of explanation-guided training on RelationNet (RN)[2], cross attention network (CAN)[3],and GNN[5] on four cross domain datasets.

miniImagenet	1-shot	1-shot-T	5-shot	5-shot-T
RN	58.31±0.47%	61.52±0.58%	72.72±0.37%	73.64±0.40%
LRP-RN	**60.06±0.47%**	**62.65±0.56%**	**73.63±0.37%**	**74.67±0.39%**
CAN	**64.66±0.48%**	67.74±0.54%	79.61±0.33%	80.34±0.35%
LRP-CAN	64.65±0.46%	**69.10±0.53%**	**80.89±0.32%**	**82.56±0.33%**
mini-CUB	1-shot	1-shot-T	5-shot	5-shot-T
RN	41.98±0.41%	42.52±0.48%	58.75±0.36%	59.10±0.42%
LRP-RN	**42.44±0.41%**	**42.88±0.48%**	**59.30±0.40%**	**59.22±0.42%**
CAN	44.91±0.41%	46.63±0.50%	63.09±0.39%	62.09±0.43%
LRP-CAN	**46.23±0.42%**	**48.35±0.52%**	**66.58±0.39%**	**66.57±0.43%**
mini-Cars	1-shot	1-shot-T	5-shot	5-shot-T
RN	29.32±0.34%	28.56±0.37%	38.91±0.38%	37.45±0.40%
LRP-RN	**29.65±0.33%**	**29.61±0.37%**	**39.19±0.38%**	**38.31±0.39%**
CAN	31.44±0.35%	30.06±0.42%	41.46±0.37%	40.17±0.40%
LRP-CAN	**32.66±0.46%**	**32.35±0.42%**	**43.86±0.38%**	**42.57±0.42%**
mini-Places	1-shot	1-shot-T	5-shot	5-shot-T
RN	**50.87±0.48%**	**53.63±0.58%**	66.47±0.41%	67.43±0.43%
LRP-RN	50.59±0.46%	53.07±0.57%	**66.90±0.40%**	**68.25±0.43%**
CAN	56.90±0.49%	60.70±0.58%	72.94±0.38%	74.44±0.41%
LRP-CAN	**56.96±0.48%**	**61.60±0.58%**	**74.91±0.37%**	**76.90±0.39%**
mini-Plantae	1-shot	1-shot-T	5-shot	5-shot-T
RN	33.53±0.36%	33.69±0.42%	47.40±0.36%	46.51±0.40%
LRP-RN	**34.80±0.37%**	**34.54±0.42%**	**48.09±0.35%**	**47.67±0.39%**
CAN	36.57±0.37%	36.69±0.42%	50.45±0.36%	48.67±0.40%
LRP-CAN	**38.23±0.45%**	**38.48±0.43%**	**53.25±0.36%**	**51.63±0.41%**

5-way 1-shot	miniImagenet	Cars	Places	CUB	Plantae
GNN	64.47±0.55%	30.97±0.37%	54.64±0.56%	46.76±0.50%	37.39±0.43%
LRP-GNN	**65.03±0.54%**	**32.78±0.39%**	**54.83±0.56%**	**48.29±0.51%**	**37.49±0.43%**
5-way 5-shot	miniImagenet	Cars	Places	CUB	Plantae
GNN	80.74±0.41%	42.59±0.42%	72.14±0.45%	63.91±0.47%	**54.52±0.44%**
LRP-GNN	**82.03±0.40%**	**46.20±0.46%**	**74.45±0.47%**	**64.44±0.48%**	54.46±0.46%

- The combination of explanation-guided training and feature-wise transformation layer[4] using RelationNet. The further improvement verify the non-overlap between LFT and explanation-guided training.

5-way 1-shot	Cars	Places	CUB	Plantae
RN	29.40±0.33%	48.05±0.46%	44.33±0.43%	34.57±0.38%
FT-RN	30.09±0.36%	48.12±0.45%	44.87±0.44%	35.53±0.39%
LRP-RN	30.00±0.32%	48.74±0.45%	45.64±0.42%	36.04±0.38%
LFT-RN	30.27±0.34%	48.07±0.46%	47.35±0.44%	35.54±0.38%
LFT-LRP-RN	**30.68±0.34%**	**50.19±0.47%**	**47.78±0.43**	**36.58±0.40%**
5-way 5-shot	Cars	Places	CUB	Plantae
RN	40.01±0.37%	64.56±0.40%	62.50±0.39%	47.58±0.37%
FT-RN	40.52±0.40%	64.92±0.40%	61.87±0.39%	48.54±0.38%
LRP-RN	41.05±0.37%	66.08±0.40%	62.71±0.39%	48.78±0.37%
LFT-RN	41.51±0.39%	65.35±0.40%	64.11±0.39%	49.29±0.38%
LFT-LRP-RN	**42.38±0.40%**	**66.23±0.40%**	**64.62±0.39%**	**50.50±0.39%**

- Qualitative LRP heatmaps .

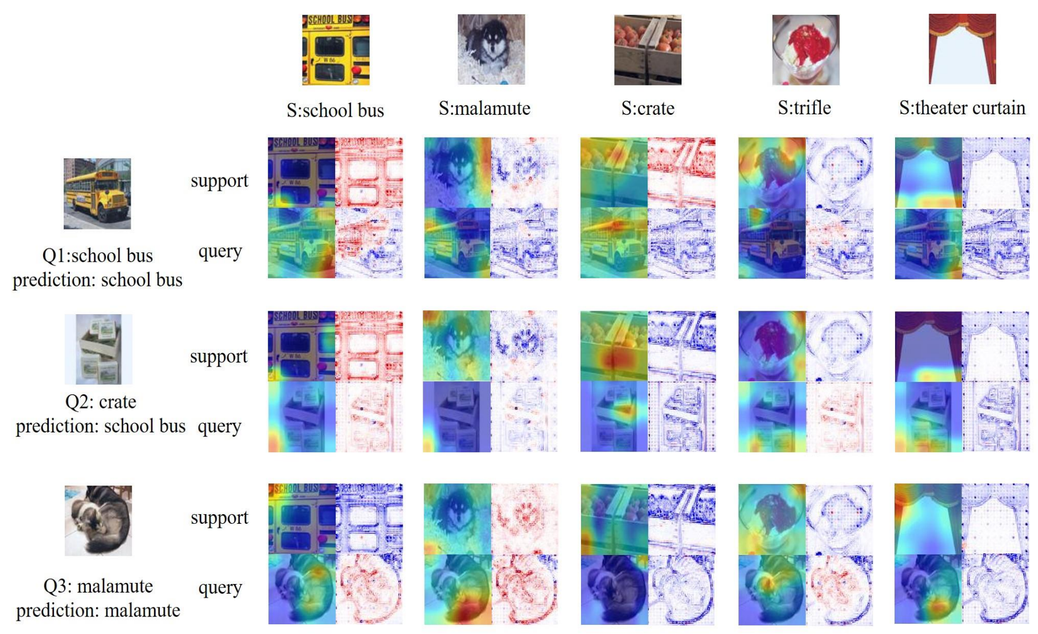

Figure 2. LRP heatmaps and attention heatmaps of the CAN model under 5-way 1-shot setting

Conclusion

- Explanation-guided training successfully addresses the domain-shift problem in few-shot learning.
- When combining explanation-guided training with learned feature-wise transformation layers, the model performance is further improved, indicating that these two approaches optimize the model in a non-overlapping manner.

References

[1]. S. Bach, A. Binder, G. Montavon, F. Klauschen, K. M¨uller, andW. Samek, “On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation,”PloS one, vol. 10, no. 7, p.e0130140, 2015
[2]. F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. Torr, and T. M. Hospedales,“Learning to compare: Relation network for few-shot learning,” inProceedings of the IEEE Conference on Computer Vision and PatternRecognition, 2018, pp. 1199–1208
[3]. Hou, H. Chang, M. Bingpeng, S. Shan, and X. Chen, “Cross attention network for few-shot classification,” in Advances in Neural Information Processing Systems, 2019, pp. 4005–4016
[4]. H.-Y. Tseng, H.-Y. Lee, J.-B. Huang, and M.-H. Yang, “Cross-domain few-shot classification via learned feature-wise transformation,” inICLR,2020
[5]. V. G. Satorras and J. B. Estrach, “Few-shot learning with graph neural networks,” in ICLR, 2018.