Explain and Improve: Cross-Domain Few-Shot-Learning Using Explanations

Jiamei Sun [1] Sebastian Lapuschkin [2] Wojciech Samek [2] Yunqing Zhao [1] Ngai-Man Cheung [1]
Alexander Binder [1]

Abstract

Cross-domain few-shot learning (CD-FSL) has attracted much interest recently. In CD-FSL, 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. In this paper, we introduce a novel approach for CD-FSL by leveraging the explanations of the FSL models. First, we tailor the layer-wise relevance propagation (LRP) method to explain the FSL models. Second, we develop a model-agnostic explanation-guided training strategy that dynamically finds and emphasizes the features that are important to the predictions. We show that, without introducing more parameters, explanation-guided training effectively improves the model generality under the cross-domain setting. We observe improved accuracy of two FSL models: RelationNet (Sung et al., 2018), and cross attention net (CAN) (Hou et al., 2019), on five few-shot learning datasets: miniImagenet, CUB, Cars, Places, and Plantae, which are introduced by (Tseng et al., 2020).

1. Introduction

In the past years, the explainability of predictors has attracted attention in the machine learning community. A large number of approaches have been developed. In this paper, we do not present another new method. Instead, we touch onto a mildly neglected dimension, namely the problem statements for which explainability is useful. Known use cases are auditing of predictions(Lapuschkin et al., 2019) and identification of biases in datasets(Selvaraju et al., 2017). In this paper, we add a different use case. We consider the question of whether explanations are suitable to improve model performance in small sample size regimes. We choose few-shot classification (FSC) as the subject of our study (Vinyals et al., 2016; Finn et al., 2017; Snell et al., 2017; Sung et al., 2018; Satorras & Estrach, 2018; Rusu et al., 2019; Sun et al., 2019; Hou et al., 2019).

Few-shot learning has two notable properties. Firstly, it aims at generalization across sets of labeled tasks. Secondly, it relies on small sample sizes. In other settings, the first choice for better generalization would be to label more training data. The question arises whether explanation scores computed for intermediate feature maps can be employed to improve generalization in few-shot learning in lieu of unavailable training data. This is not assured, as explanations are computed on a per-sample basis, and, unlike data-augmentation techniques do not create additional samples.

Commonly, FSC models are evaluated using a test dataset originating from the same domain as the training dataset. (Chen et al., 2019) states that FSC methods will meet difficulties in cases with domain shift between the training data (source domain) and the test data (target domain). To tackle the domain shift problem, we need to avoid overfitting to the source domain. A recent work achieves this by learning a noise distribution for some intermediate layers in the feature encoder (Tseng et al., 2020), while others rely on adding batch spectral regularization over the encoded image features (Liu et al., 2020), and employing novel losses (Yeh et al., 2020; Chen et al., 2019). We propose an approach from a different perspective: we leverage the explanations for FSC methods to guide the model to learn features with better performance.

We adapt LRP-type explanations(Bach et al., 2015) to FSC models. LRP has generated usable explanations for CNN (Bach et al., 2015), RNN(Arras et al., 2017), graph neural networks(GNN)(Schnake et al., 2020), and clustering models(Kauffmann et al., 2019). It backpropagates a score through the neural network and assigns relevance scores to the neurons within the network. The LRP scores reflect the importance of a neuron to the prediction, which we can easily observe in Figure 1. Relying on this property, we propose explanation-guided training for FSC models. The LRP relevance scores of intermediate features are employed as weights. We construct the LRP-weighted feature

[1] Information Systems Technology and Design Pillar, Singapore University of Technology and Design, Singapore [2] Department of Video Coding & Analytics, Fraunhofer Heinrich Hertz Institute, Berlin, Germany. Correspondence to: Jiamei Sun <jiamei_sun@mymail.sutd.edu.sg>.

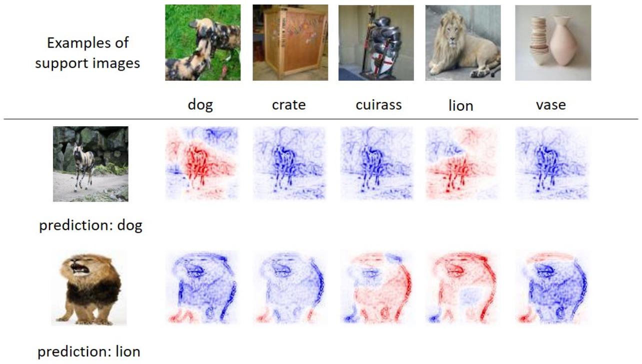

Figure 1. LRP explanation heatmaps of the input image with 5 target labels. The experiment model is a RelationNet trained on miniImagenet under the 5-way 5-shot setting. The first row illustrates some example support images. The other two rows show the explanation heatmaps of two query images, the *African hunting dog* (denoted as *dog* in the figure) and the *lion*. Both images are correctly predicted and the heatmaps are generated using different target labels. Red pixels indicate positive LRP relevance score and blue indicates negative. The strength of the color corresponds to the value of the LRP relevance scores.

maps, which emphasize features that are more relevant to the model predictions and downscale less relevant features. LRP explanations are calculated for each sample-label pair separately. The LRP-weighted features are fed into the network to guide the training. This adds a label-dependent feature weighting mechanism during training, reducing overfitting to the source domain.

The main contributions of this paper are 1) We derive explanations for FSC models using LRP; 2) We investigate the potential for improvement of model performance using explanations in the training phase under small sample size settings. 3) We propose an explanation-guided training strategy to tackle the domain shift problem in FSC; We remark that the principles used for explanation-guided training strategy are model-agnostic and can be combined with many other methods such as learned feature-wise transformation (LFT) (Tseng et al., 2020). We will verify that LRP-weighted features during training are improving generalization.

2. Related Work

Cross-domain few-shot classification methods It is common to develop cross-domain few-shot classification methods by basing them on existing FSC methods. LFT(Tseng et al., 2020) learns a noise distribution and adds the noise on some intermediate feature maps to generate more diverse features during training and improve the model generality. In the most recent CVPR Cross-Domain Few-Shot Learning challenge, (Liu et al., 2020) ensembled multiple feature encoders and employed batch spectral regularization over the image features for each encoder. Batch spectral regularization penalizes the singular values of the feature matrix within a batch so that the learned features maintain similar spectra across domains. (Cai & Shen, 2020) combined the first-order MAML(Finn et al., 2017) and the GNN metric-based method (Satorras & Estrach, 2018). (Yeh et al., 2020) applied prototypical triplet loss to increase the inter-class distance and large margin cosine loss to minimize the intra-class distance, which is also studied by (Chen et al., 2019) that reducing intra-class variation benefits FSC especially for shallow image feature encoders.

In our approach, we do not introduce more parameters like (Tseng et al., 2020). We are similar to (Liu et al., 2020; Yeh et al., 2020) in adding constraints on the image features. We use LRP-weighted features to guide the model to dynamically correct itself for each instance instead of penalizing feature statistics over a batch.

Explanation for few-shot classification models To our best knowledge, there have not been explanation methods specially designed for FSC models. On the other hand, there exist explanation methods for deep neural networks(DNN) (Bach et al., 2015; Montavon et al., 2017; Simonyan et al., 2013; Springenberg et al., 2015; Selvaraju et al., 2017; Schnake et al., 2020) which can be adapted to FSC models, since many FSC models adopt CNN to encode image features and many metric-based methods also adopts DNN to learn the distance metric (Sung et al., 2018; Satorras & Estrach, 2018; Liu et al., 2019). For FSC models that use non-parametric distance metrics, we refer to (Kauffmann et al., 2019) which neuralizes various K-means classifiers and applies LRP to obtain explanations.

In this paper, we choose LRP to explain FSC models due to its general applicability on both parametric and non-parametric classifiers.

3. Explanation-guided training.

Before presenting our explanation-guided training, we first introduce the cross-domain few-shot learning task and some notations. For a K-way N-shot task, denoted as an episode, we are given a support set $\mathcal{S} = \{(x_s, y_s)\}_{s=1}^{K*N}$ containing K classes and N labeled samples per class for training and a query set $\mathcal{Q} = \{(x_q, y_q)\}_{q=1}^{n_q}$ from the same classes as $\mathcal{S}$ for testing. A CD-FSC task is to train an FSC model using episodes $\{\mathcal{S}_i, \mathcal{Q}_i\}$ randomly sampled from a base domain $\mathcal{D}_{seen}$ and test the model with episodes sampled from an unseen domain $\mathcal{D}_{unseen}$. We consider FSC models which can be summarized as in Figure 2. This includes a number of metric-based FSC models.

The support set $\mathcal{S}$ and query set $\mathcal{Q}$ are encoded by a CNN (Sung et al., 2018; Hou et al., 2019), possibly with augmented layers (Oreshkin et al., 2018; Tseng et al., 2020) to

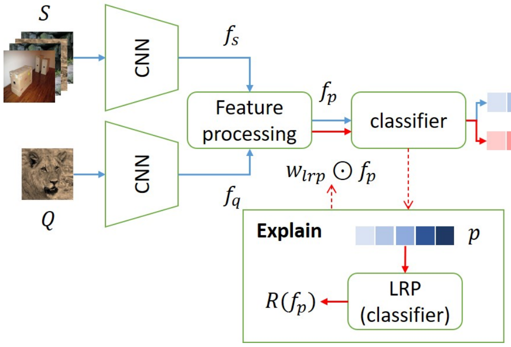

Figure 2. Explanation-guided training. Blue paths denote the conventional FSC training. The red paths are added after one step relying on the blue paths. The support samples $\mathcal{S}$ and the query sample $\mathcal{Q}$ are fed into an image encoder to obtain features f_s and f_q, which are compared by a *feature processing* module. The output of *feature processing* f_p is fed into a classifier to make predictions. Both the *feature processing* and *classifier* modules vary across different FSC methods. The **Explain** block explains the model prediction p and generate the explanations for f_p, denoted as $R(f_p)$, which are used to calculate the LRP weight w_{lrp}. This is fed into the classifier resulting in the updated prediction p_{lrp}

obtain the support image features f_s and the query image features f_q. f_s and f_q are further processed before classification, for example, (Sung et al., 2018) simply averages the f_s over classes and concatenate the averaged class representations pairwise with f_q, (Hou et al., 2019) designs an attention module and generate the attention-weighted support and query image features, (Liu et al., 2019) applies GNN on f_s and f_q to obtain graph structured features.

The processed features are fed into a classifier for generating predictions. The classifier can be cosine a similarity(Hou et al., 2019), Euclidean distances(Snell et al., 2017), or neural nets(Sung et al., 2018; Satorras & Estrach, 2018).

Explanation-guided training for FSC models involves the following steps. For each training episode:

Step1: One forward-pass through the model and obtain the prediction p, illustrated as the blue path in Figure 2.

Step2: **Explaining the classifier**. We initialize the LRP relevance for each label and apply LRP to explain the classifier. We can obtain the relevance of the classifier input $R(f_p)$, illustrated as the **Explain** block. For FSC models which implement a neural network as the classifier, the relevance scores for each label can be initialized with their logits. For the models using non-parametric distance measures such as cosine similarity and Euclidean distance, the predicted scores are positive for all labels, which will result in similar explanations. For such cases, we refer to the logit function in (Kauffmann et al., 2019) to initialize the relevance score. Taking the cosine similarity as an example, we first calculate the probability for each class using the exponential function as eq(1) [1].

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

$cs_k(\cdot)$ means the cosine similarity between a query sample and class k. f_p is the processed feature fed to the classifier. β is a constant scale parameter to strengthen the highest probability. With the probability, the relevance score of class c is defined as:

$$R_c = log\left(\frac{P(y_c|f_p)}{1 - P(y_c|f_p)}(K-1)\right) \tag{2}$$

R_c is positive when the $P(y_c|f_p)$ is larger than $1/(K)$. In other words, the class label whose probability is larger than the random guessing probability receives a positive relevance score.

With the relevance score of each target label $R_c, c = 1 \dots K$, standard LRP is applicable to backpropagate R_c through the classifier to generate the explanations. We rely on two established LRP backpropagation mechanisms here, the ϵ-rule and the $\alpha\beta$-rule (Bach et al., 2015). Consider the forward pass from layer l to $l+1$: $y_j^{l+1} = \sum_i w_{ij} z_i^l + b_j, z_j^{l+1} = f(y_j^{l+1})$ where i and j are the indices of neurons in l^{th} and $l+1^{th}$ layer, $f(\cdot)$ is an activation function. Let $R(\cdot)$ denote the relevance of a neuron and $R_{i \leftarrow j}$ denote the relevance attribution from z_j^{l+1} to z_i^l.

ϵ**-rule**: $R_{i \leftarrow j} = R(z_j^{l+1}) \frac{z_i^l w_{ij}}{y_j^{l+1} + \epsilon \text{sign}(y_j^{l+1})}$. ϵ is a small positive number and $\epsilon\text{sign}(y_j^{l+1})$ guarantees safe division.

$\alpha\beta$**-rule**:$R_{i \leftarrow j} = R(z_j^{l+1})(\alpha \frac{(z_i^l w_{ij})^+}{(y_j^{l+1})^+} - (\alpha - 1) \frac{(z_i^l w_{ij})^-}{(y_j^{l+1})^-})$ where $(y_j^{l+1})^+ = \max(y_j^{l+1}, 0)$ and $\alpha \geqslant 1$ controls the ratio of positive evidence to backpropagate.

The relevance of z_i^l is the summation of the relevance attribution of it, $R(z_i^l) = \sum_j R_{i \leftarrow j}$. We adopt the ϵ-rule and $\alpha\beta$-rule for linear layer and convolutional layer respectively to obtain $R(f_p)$. $R(f_p)$ is normalized with its maximal absolute value.

Step3: **LRP-weighted features**. To emphasize features which are more relevant to the prediction and downscale the less relevant ones, we define the LRP weights and the LRP-weighted features as

$$w_{lrp} = 1 + R(f_p) \tag{3}$$

$$f_{p-lrp} = w_{lrp} \odot f_p \tag{4}$$

Note that $R(f_p) \in [-1, 1]$ after normalization, thus w_{lrp} magnifies the features with positive relevance scores and

[1] For Euclidean distance, we need to use the opposite number of the distance to replace the similarity metric.

Table 1. The correspondence of RelationNet and Cross attention network to the framework in Figure 2.

	feature processing	classifier
RN	pairwise concatenation	relation module
CAN	cross attention module	cosine similarity

downscales those with negative relevance scores. The maximal feature scaling after weighting with w_{lrp} is 2.

Step4: Finally, we forward the LRP-weighted features to the classifier to generate the explanation-guided predictions p_{lrp}. The objective function merges both the model prediction p and the explanation-guided prediction p_{lrp}.

$$\mathcal{L} = \xi \mathcal{L}_{ce}(y, p) + \lambda \mathcal{L}_{ce}(y, p_{lrp}) \tag{5}$$

where $\mathcal{L}_{ce}$ is the cross entropy loss. ξ and λ are positive scalars that control how much information from p and p_{lrp} are used. In our experiment, ξ and λ are empirically adjusted for the different FSC models.

4. Experiments

We evaluate the proposed explanation-guided training on RelationNet(RN)(Sung et al., 2018) and one of the state-of-the-art models, cross attention network(CAN) (Hou et al., 2019). The correspondence of the two FSC models to the framework in Figure 2 is summarized in Table 1. We will prove that explanation-guided training improves the prediction performance of both RN and CAN on 4 cross-domain test sets.

Moreover, we also combine explanation-guided training with another approach, the learning to learn feature-wise transformation (LFT)(Tseng et al., 2020). We show that explanation-guided training is compatible with LFT and the combination further improves the performance.

4.1. Dataset and model preparation

Five datasets are used in our experiment including miniImagenet(Ravi & Larochelle, 2016), CUB(Wah et al., 2011), Cars(Krause et al., 2013), Places(Zhou et al., 2017), and Plantae(Van Horn et al., 2018), which are introduced in (Tseng et al., 2020). Each dataset consists of train/val/test splits. We choose miniImagenet as the $\mathcal{D}_{seen}$ and train the RN and CAN models on the training set, validate both models on the validation set of miniImagenet, and adopt the test sets of the other four datasets for testing.

We use Resnet10 and Resnet12(He et al., 2016) as the image encoder for both RN and CAN models respectively. The two models are trained under 5-way 5-shot and 5-way 1-shot settings. For explanation-guided training, we set $\xi = 1, \lambda = 0.5$ for RN 5-way 5-shot setting and $\xi = 1, \lambda = 1$ for RN 5-way 1-shot setting. CAN model employs cosine similarity as the classifier, thus we set β in eq(1) as 7, the same as the original model and $\xi = 0, \lambda = 1$ for eq(5). The LRP parameters are $\alpha = 1, \epsilon = 0.001$ for all the experiments.

We follow the same implementation details (Tseng et al., 2020)[2] and (Hou et al., 2019)[3] to train the RelationNet and CAN model. At test time, we evaluate the performance over 2000 randomly sampled episodes, with 16 query images per episode.

4.2. Evaluation for explanation-guided training on cross-domain setting

In this section, we evaluate the performance of RN and CAN models trained without and with explanation-guided training on CD-FSC tasks. For more comprehensive analyses, we also implement the **Transductive inference** proposed by (Hou et al., 2019). Transductive inference iteratively augments the support set using the confidently classified query images during the test phase. Specifically, we first predict the label of query images with the trained model; secondly, we choose the query images with higher predicted scores as the candidate images. The candidate images and their predicted label are augmented to the support set. This is an iterative process. In our experiment, we implement the transductive operation for two iterations with 35 candidates for the first iteration and 70 for the second which is the same strategy as (Hou et al., 2019).

Table 2 summarizes the accuracy of the unmodified RN and CAN models and the same models with explanation-guided training. We can observe a consistent improvement after implementing explanation-guided training. The results are also competitive with the recent work on LFT (Tseng et al., 2020) which learns a noise distribution by adding feature-wise transformation layers to the image encoder while explanation-guided training does not introduce more training parameters. To show that our approach exploits a different mechanism to improve performance, we combine the LFT and our explanation-guided training.

4.3. Collaboration of explanation-guided training and feature-wise transformation

To compare and to combine our idea with the LFT method, we apply the explanation-guided training to the multiple domain experiment as (Tseng et al., 2020). The LFT model is trained using the pseudo-seen domain and pseudo-unseen domains. In our experiment, the miniImagenet is the pseudo-seen domain. Three of the other four datasets are the pseudo-unseen domains and the model is tested on the last domain. The pseudo-unseen domains are used to train the feature-

[2] https://github.com/hytseng0509/CrossDomainFewShot

[3] https://github.com/blue-blue272/fewshot-CAN

Table 2. Evaluation of explanation-guided training on cross-domain datasets using RN and CAN. We report the average accuracy of over 2000 episodes with 95% confidence intervals. The models are trained on the miniImagenet training set and tested on the test set of various domains. **LRP-** means explanation-guided training using LRP. **T** indicates transductive inference.

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**%

Table 3. The results of multiple domains experiment under the 5-way 5-shot setting. We report the average accuracy of over 2000 episodes with 95% confidence intervals. **FT** and **LFT** indicate the feature-wise transformation layer with fixed or trainable parameters. **LRP-** means explanation-guided training using LRP. **LFT-LRP** is the combination of LFT and explanation-guided training.

	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**%

wise transformation layer and the pseudo-seen domain is used to update the other trainable parameters of the model. If the parameters of the feature-wise transformation layer are fixed, we will get the FT method that adds the noise with a fixed distribution on certain intermediate layers.

The performance of the standard RN, the FT and LFT methods, explanation-guided training, and its combination with LFT are shown in Table 3. These models are trained with the same random seed, learning rate, optimizer, and datasets. The combination of our explanation-guided training and LFT(**LFT-LRP-RN**) achieves the best accuracy. Comparing the results of **FT-RN** and **LRP-RN**, we can see explanation-guided training is even better without introducing more trainable parameters to the model.

We remark that the improvement observed when combining LRP with LFT shows that both optimize the model from different angles. This demonstrates the independence of both approaches as well as their strength.

4.4. Qualitative results of LRP explanation for FSC models

The above experiments have demonstrated that explanation-guided training effectively improves the performances of FSC models and successfully reduces the domain gap. We leverage on the LRP explanation of the intermediate feature map to re-weight the same feature map. In this section, we visualize the LRP explanation of the input images as heatmaps. From the LRP heatmaps, we can easily observe which parts of the image are used by the model to make the predictions, in other words, what features have the model learned to differentiate classes. To our best knowledge, this is the first attempt to explain the FSC models though many existing explanation methods are in principle applicable.

Figure 1 has already shown some heatmaps of RelationNet.

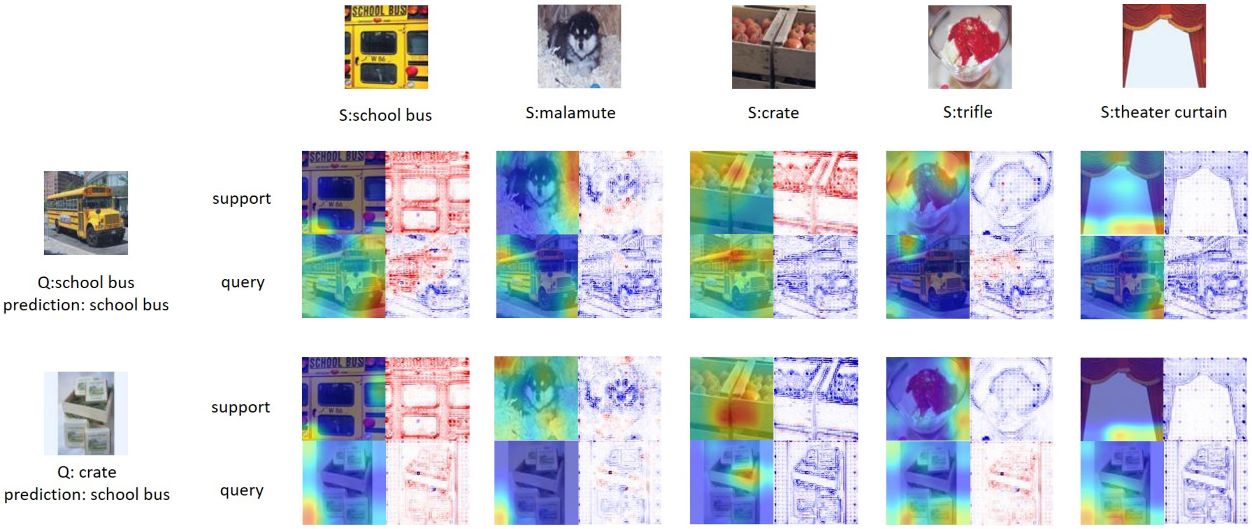

Figure 3. LRP heatmaps and the attention heatmaps of the CAN model for one episode. The model is trained under the 5-way 1-shot setting. The first row shows the support images of each class. For each query image, we illustrate the attention heatmaps and the LRP heatmaps of both the support images and the query images with 5 target labels.

We further illustrate the LRP explanation of the CAN model in the 5-way 1-shot setting. Since there is only one training sample for each class, we also show the LRP heatmap and the attention heatmap for the support images.

In Figure 3, we can see the LRP heatmaps and attention heatmaps of each label for the query image. For the query image that is correctly predicted as *school bus*, LRP heatmaps under *school bus* highlight the relevant structures of the bus. Specifically, the LRP heatmap can capture the features of the window frames of the bus. On the other hand, the LRP heatmaps of other wrong labels show more negative evidence, however, we can still find some interesting resemblance between the query image and the explained label. For example, in Figure 1, when we explain the label *lion* for the *African hunting dog*, the LRP heatmap highlights the legs of the *African hunting dog* and when we explain the label *cuirass* for the *lion*, the LRP heatmap emphasizes the round contour that resembles cuirass.

Moreover, LRP heatmaps provide some evidence for us to analyze the reasons why the model makes wrong predictions, such as the *crate* that is wrongly predicted as *school bus* in Figure 3. The support image *school bus* contains the window frames with latticed shape which is also an obvious feature of the *crate* class, usually shown as a pile of rectangles. These features are highlighted by LRP heatmaps and we can speculate that the model perhaps makes the wrong prediction according to the similar features between the two classes.

5. Conclusion

This paper tailors LRP to explain few-shot classification models and propose a novel approach to improve FSC models, explanation-guided training. We find two points noteworthy. Firstly, explanation-guided training successfully addresses the domain shift problem in few-shot learning, as demonstrated in the cross-domain few shot tasks. Secondly , when combining explanation-guided training with feature-wise transformation, the model performance is further improved, demonstrating that these two approaches optimize the model in a non-overlapping manner. We conclude that the explanations of the few-shot classification can not only provide intuitive and informative visualizations but can also be leveraged to improve the models.

References

Arras, L., Montavon, G., Müller, K., and Samek, W. Explaining recurrent neural network predictions in sentiment analysis. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pp. 159–168, 2017.

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K., and Samek, W. On pixel-wise explanations for nonlinear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015.

Cai, J. and Shen, S. M. Cross-domain few-shot learning with meta fine-tuning. *arXiv preprint arXiv:2005.10544*, 2020.

Chen, W.-Y., Liu, Y.-C., Kira, Z., Wang, Y.-C. F., and

Huang, J.-B. A closer look at few-shot classification. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=HkxLXnAcFQ.

Finn, C., Abbeel, P., and Levine, S. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 1126–1135. JMLR. org, 2017.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Hou, R., Chang, H., Bingpeng, M., Shan, S., and Chen, X. Cross attention network for few-shot classification. In *Advances in Neural Information Processing Systems*, pp. 4005–4016, 2019.

Kauffmann, J., Esders, M., Montavon, G., Samek, W., and Müller, K.-R. From clustering to cluster explanations via neural networks. *arXiv preprint arXiv:1906.07633*, 2019.

Krause, J., Stark, M., Deng, J., and Fei-Fei, L. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pp. 554–561, 2013.

Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., and Müller, K. Unmasking clever hans predictors and assessing what machines really learn. *Nature Communications*, 10(1):1096, 2019. ISSN 2041-1723. doi: 10.1038/s41467-019-08987-4. URL https://doi.org/10.1038/s41467-019-08987-4.

Liu, B., Zhao, Z., Li, Z., Jiang, J., Guo, Y., Shen, H., and Ye, J. Feature transformation ensemble model with batch spectral regularization for cross-domain few-shot classification. *arXiv preprint arXiv:2005.08463*, 2020.

Liu, Y., Lee, J., Park, M., Kim, S., Yang, E., Hwang, S., and Yang, Y. LEARNING TO PROPAGATE LABELS: TRANSDUCTIVE PROPAGATION NETWORK FOR FEW-SHOT LEARNING. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=SyVuRiC5K7.

Montavon, G., Lapuschkin, S., Binder, A., Samek, W., and Müller, K. Explaining nonlinear classification decisions with deep taylor decomposition. *Pattern Recognition*, 65: 211–222, 2017.

Oreshkin, B., López, P. R., and Lacoste, A. Tadam: Task dependent adaptive metric for improved few-shot learning. In *Advances in Neural Information Processing Systems*, pp. 721–731, 2018.

Ravi, S. and Larochelle, H. Optimization as a model for few-shot learning. 2016.

Rusu, A. A., Rao, D., Sygnowski, J., Vinyals, O., Pascanu, R., Osindero, S., and Hadsell, R. Meta-learning with latent embedding optimization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=BJgklhAcK7.

Satorras, V. G. and Estrach, J. B. Few-shot learning with graph neural networks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=BJj6qGbRW.

Schnake, T., Eberle, O., Lederer, J., Nakajima, S., Schütt, K. T., Müller, K.-R., and Montavon, G. Xai for graphs: Explaining graph neural network predictions by identifying relevant walks. *arXiv preprint arXiv:2006.03589*, 2020.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

Simonyan, K., Vedaldi, A., and Zisserman, A. Deep inside convolutional networks: Visualising image classification models and saliency maps. *CoRR*, abs/1312.6034, 2013.

Snell, J., Swersky, K., and Zemel, R. Prototypical networks for few-shot learning. In *Advances in neural information processing systems*, pp. 4077–4087, 2017.

Springenberg, J., Dosovitskiy, A., Brox, T., and Riedmiller, M. Striving for simplicity: The all convolutional net. In *International Conference on Learning Representations (workshop track)*, 2015.

Sun, Q., Liu, Y., Chua, T.-S., and Schiele, B. Meta-transfer learning for few-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 403–412, 2019.

Sung, F., Yang, Y., Zhang, L., Xiang, T., Torr, P. H., and Hospedales, T. M. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1199–1208, 2018.

Tseng, H.-Y., Lee, H.-Y., Huang, J.-B., and Yang, M.-H. Cross-domain few-shot classification via learned feature-wise transformation. In *ICLR*, 2020.

Van Horn, G., Mac Aodha, O., Song, Y., Cui, Y., Sun, C., Shepard, A., Adam, H., Perona, P., and Belongie, S. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8769–8778, 2018.

Vinyals, O., Blundell, C., Lillicrap, T., Wierstra, D., et al. Matching networks for one shot learning. In *Advances in neural information processing systems*, pp. 3630–3638, 2016.

Wah, C., Branson, S., Welinder, P., Perona, P., and Belongie, S. The caltech-ucsd birds-200-2011 dataset. 2011.

Yeh, J.-F., Lee, H.-Y., Tsai, B.-C., Chen, Y.-R., Huang, P.-C., and Hsu, W. H. Large margin mechanism and pseudo query set on cross-domain few-shot learning. *arXiv preprint arXiv:2005.09218*, 2020.

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., and Torralba, A. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017.