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

Niklas Baack, Jan Malte Töpperwien, Peer Bastian Duensing
Poster Presentation in context of Interpretable Machine Learning

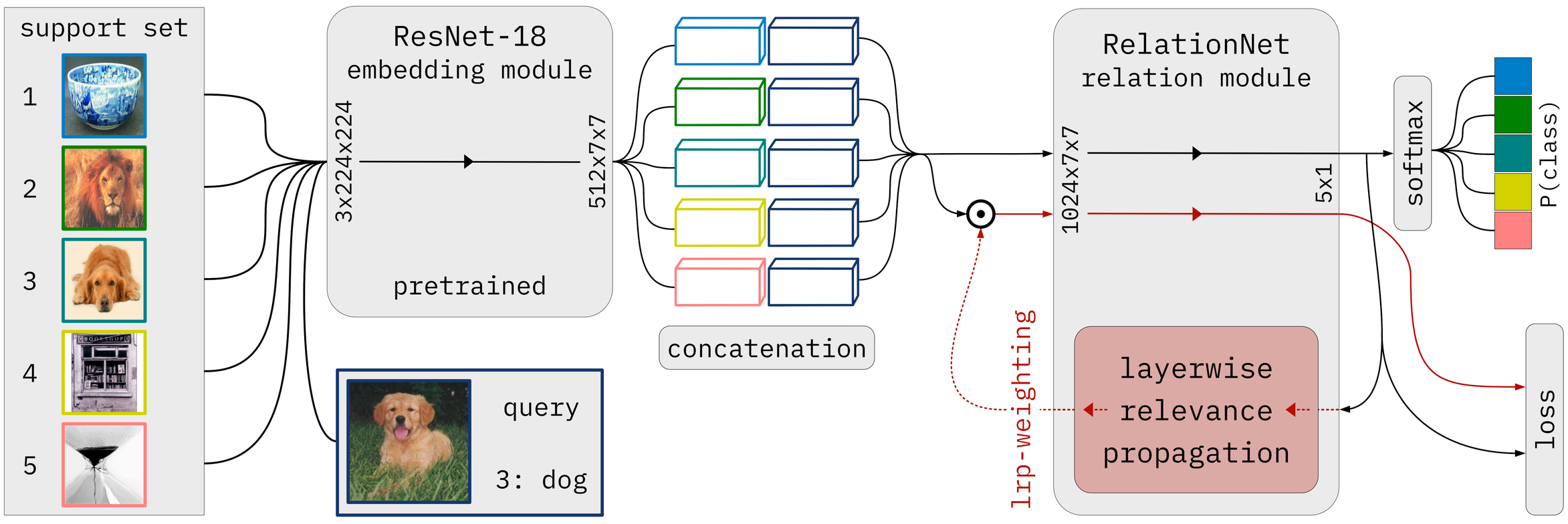
1 Summary

- LRP for guided training
- Better performance on cross-domain tasks
- Mostly model agnostic (gradients needed)

2 Motivation & Problem Setting

- Humans can identify objects with help of few examples
- Models struggle with cross-domain few-shot classification
- Utilize LRP explanations for performance improvement

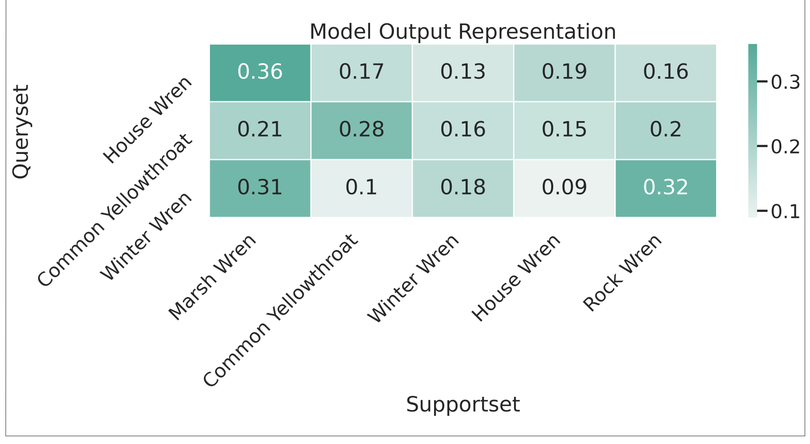
3 Approach



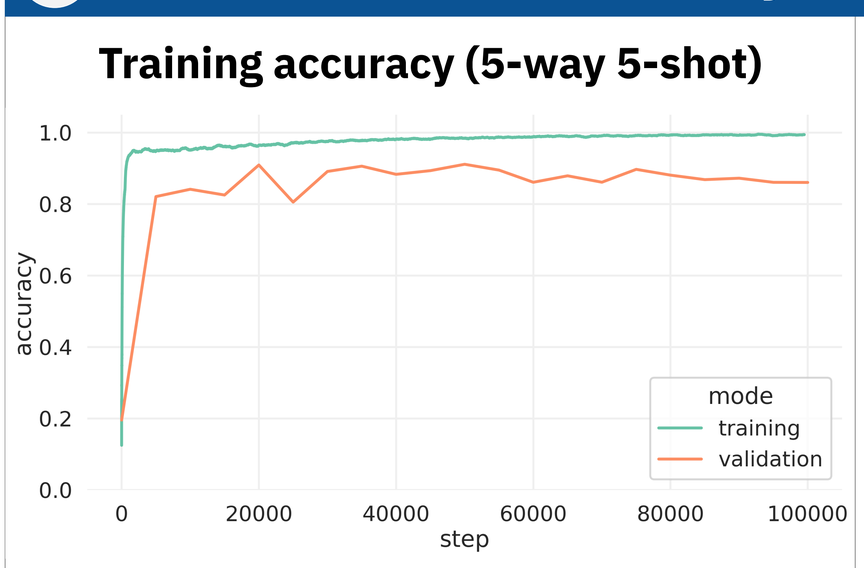
LRP rules for interpretable backpropagation

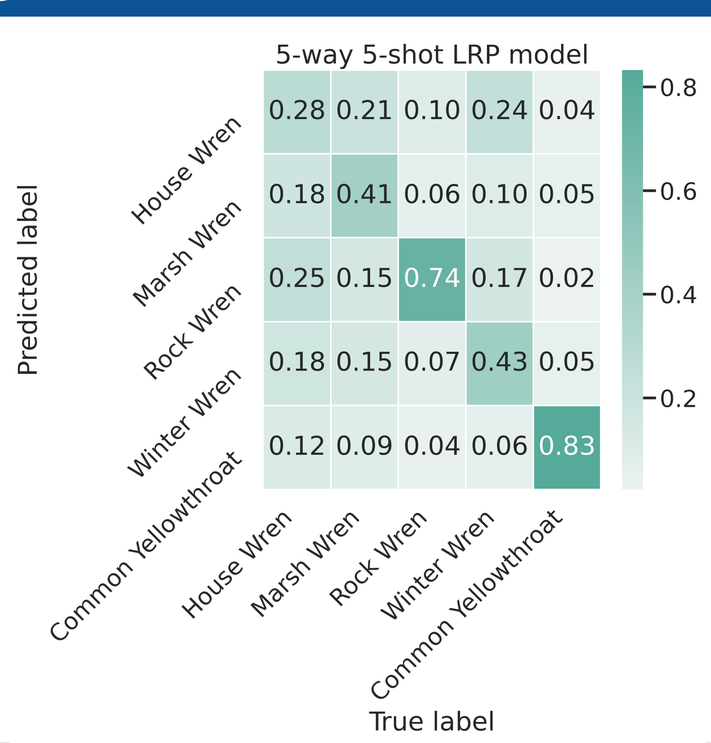
- α-rule for convolution layers
- ε-rule for linear layers
- Variation: γ-rule

$$\underbrace{R_{i\leftarrow j}}_{\text{relevance}} = R(z_j^{l+1}) \frac{\overbrace{z_i^l w_{ij}}^{\text{gradient}}}{\underbrace{\sum_i z_i^l w_{ij} + b_j}_{\text{layer normalization}}}$$
(upstream relevance)

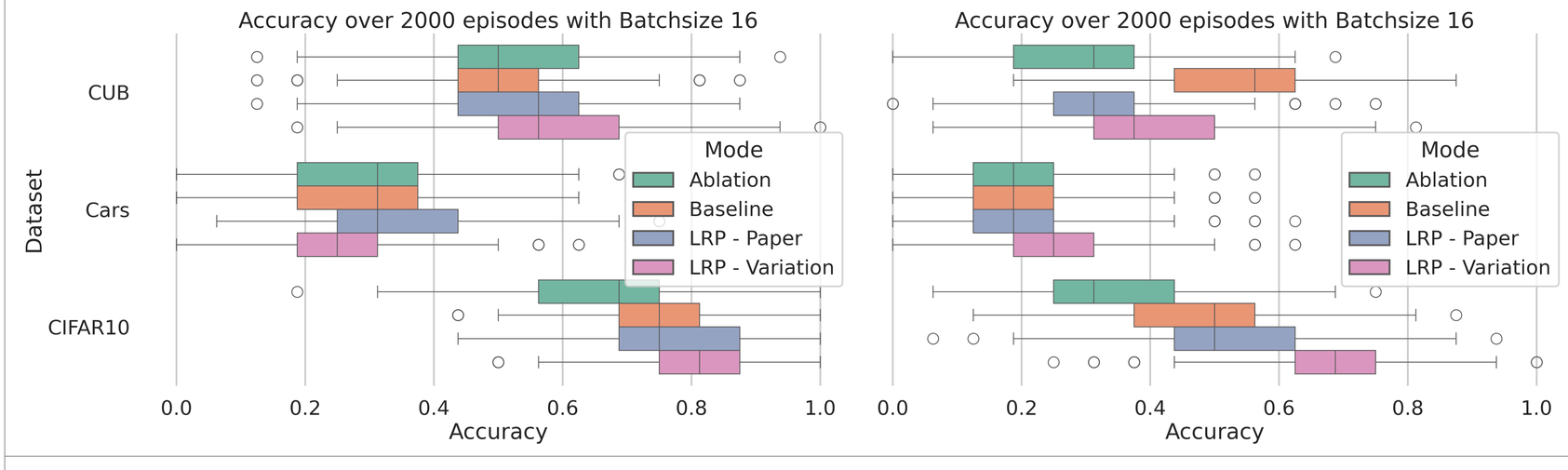


4 Key Insights





Cross-Domain Classification



- Advanced generalization by noise surpression (LRP rules)
- Better embeddings boost performance significantly

5 Future Works

- Apply on other architectures
- Use on harder domains
- Use with more classes