On the Study of Explainable AI of Few Shot Meta Learning and CNN

Hypothesis: While a fully connected CNN (Resnet 34) is obviously a better predictor than a few-shot learner if the available dataset is provided, we believe that a few-shot classifier might provide better explanations, because the model actually has to learn the logic to transfer.

Methodology: To provide a comparison, we trained a Few-Shot Meta-Learner and a ResNet34 on the CalTech256 dataset in PyTorch. I trained a 5-way-5-Shot Learner on 80% of the classes, and evaluated it on the rest of the classes. I then tested explanations of the model on the following classes: electric-guitar, zebra, tweezer, vcr, and yo-yo. I trained the ResNet only on the abovementioned 5 classes. We then provide some explanations using Local-Interpretable Model-Agnostic Explanation.

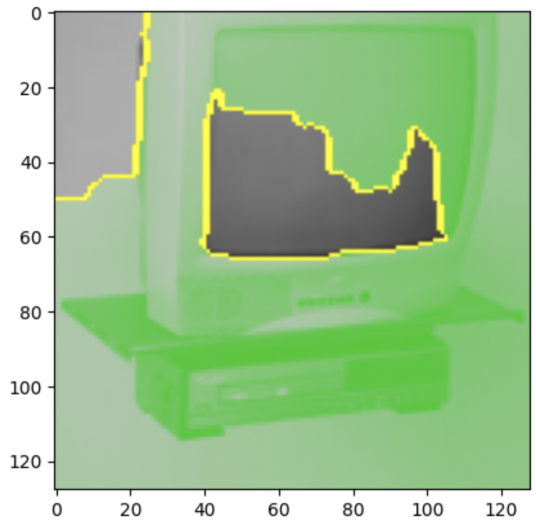
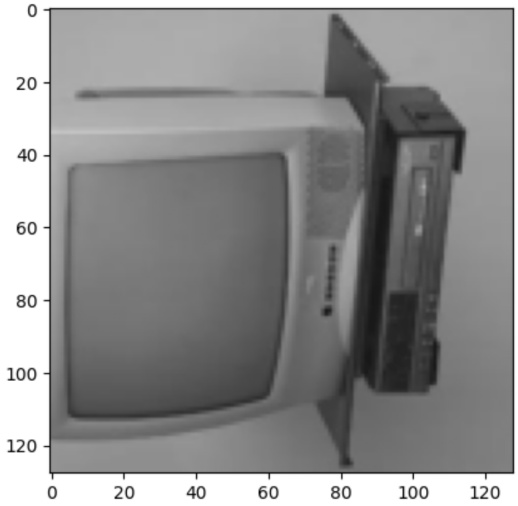
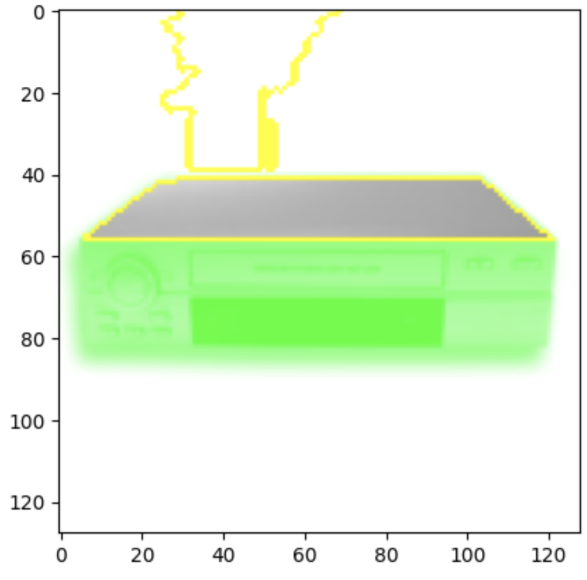
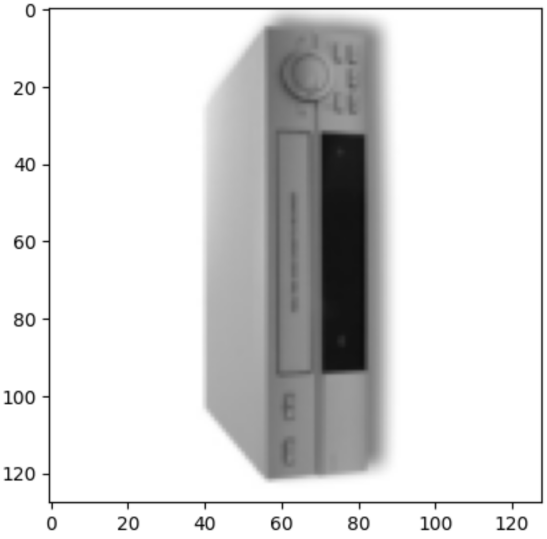
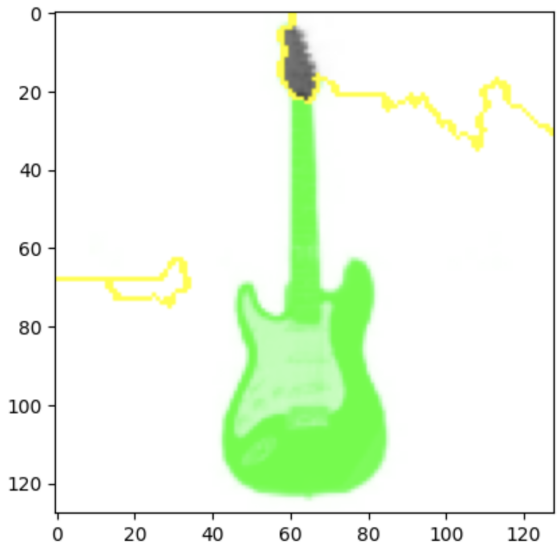
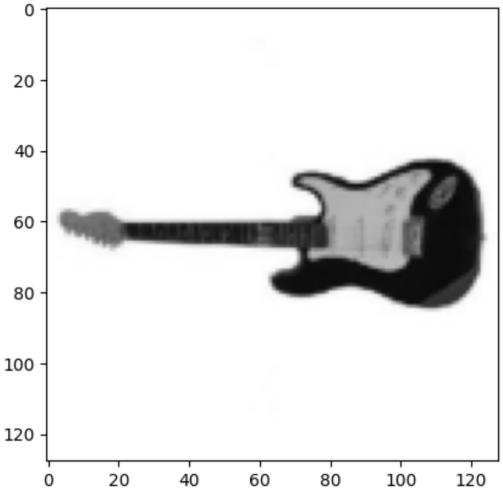
Conclusions:

Possible problems

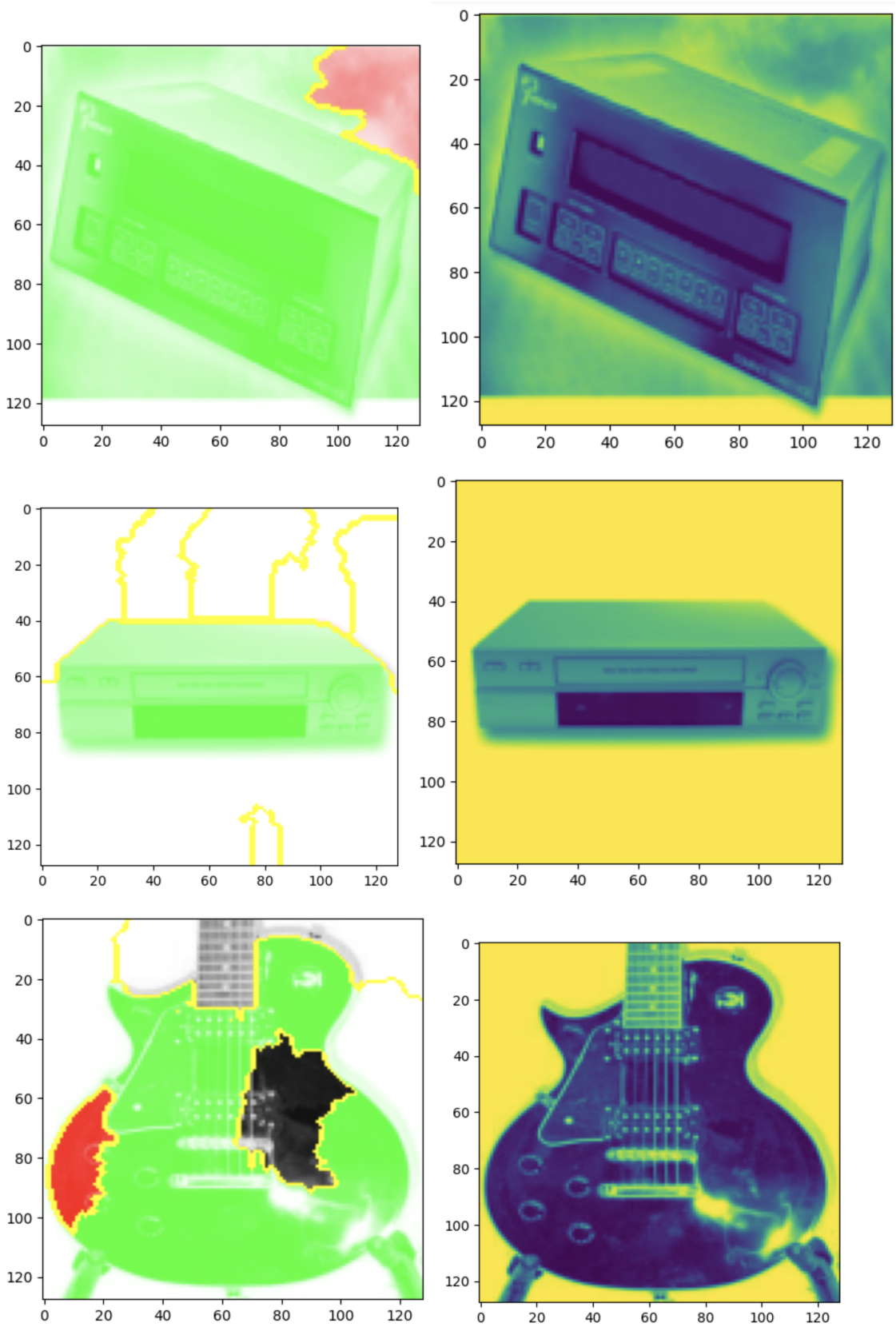
One clear possible issue is that LIME is a blackbox explainer. Therefore, even if the Meta-Learner is able to capture some intrinsic logic, LIME may not be able to pick it up because it is a blackbox learner. Second of all, the metric is not too comparable because the few-shot learner achieved an accuracy of 70% while the ResNet32 had a 90% accuracy. Therefore, the two models are not exactly comparable. Furthermore, because LIME does not provide a "explanibility" metric, our methodology is not very objective, relying on a human to give a subjective answer for "better explained" or "not as well explained". Due to the way we randomized the train/test split and conduct preprocessing (ResNet splits the data all equally, whereas few-shot splits the data by class, where 80% of the classes are all in training dataset and 20% are all in testing dataset), it is difficult to pick the same images and demo them side by side.

Results

Some explanations from ResNet34:

Some explanations from Few-Shot Learner:



In the Jupiter notebook, by changing the “sample_image_index” on both notebooks, you can generate more demo test cases. Although it is similar, it is pretty obvious that ResNet is better in terms of explanability compared to Few-Shot Explainer, although the difference is not as stark as you would expect for a 20% difference in accuracy.

Conclusions

Based on these demos (and others seen during our trials, which can be conducted through the notebooks conducted), it is pretty clear that our hypothesis is not substantiated: ResNet provides better explainability compared to Few-Shot Classification.

On all three images, ResNet is able to provide (in my mind) a pretty prediction. However, for few-shot learning, in the first demo, we see that few-shot predicts using the whole image, and in the third image of the few-shot demo, we see that the whole guitar is not covered.

Future Work

- Making the process more objective (comparing images side-by-side) would be very helpful.
- Using a gray-box method might be a lot more interesting in terms of investigating the few-shot meta-learning algorithm and how it “learns to learn”. In hindsight, it was counterproductive to use a black-box method when the whole point is to investigate few-shot learning’s idea of “learning to learn”.