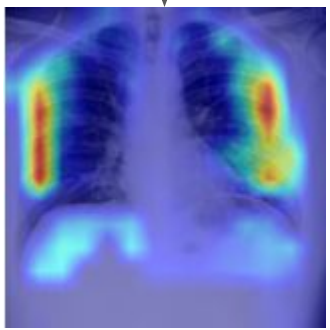
Vision transformer medical imaging explanations using Relevance Propagation

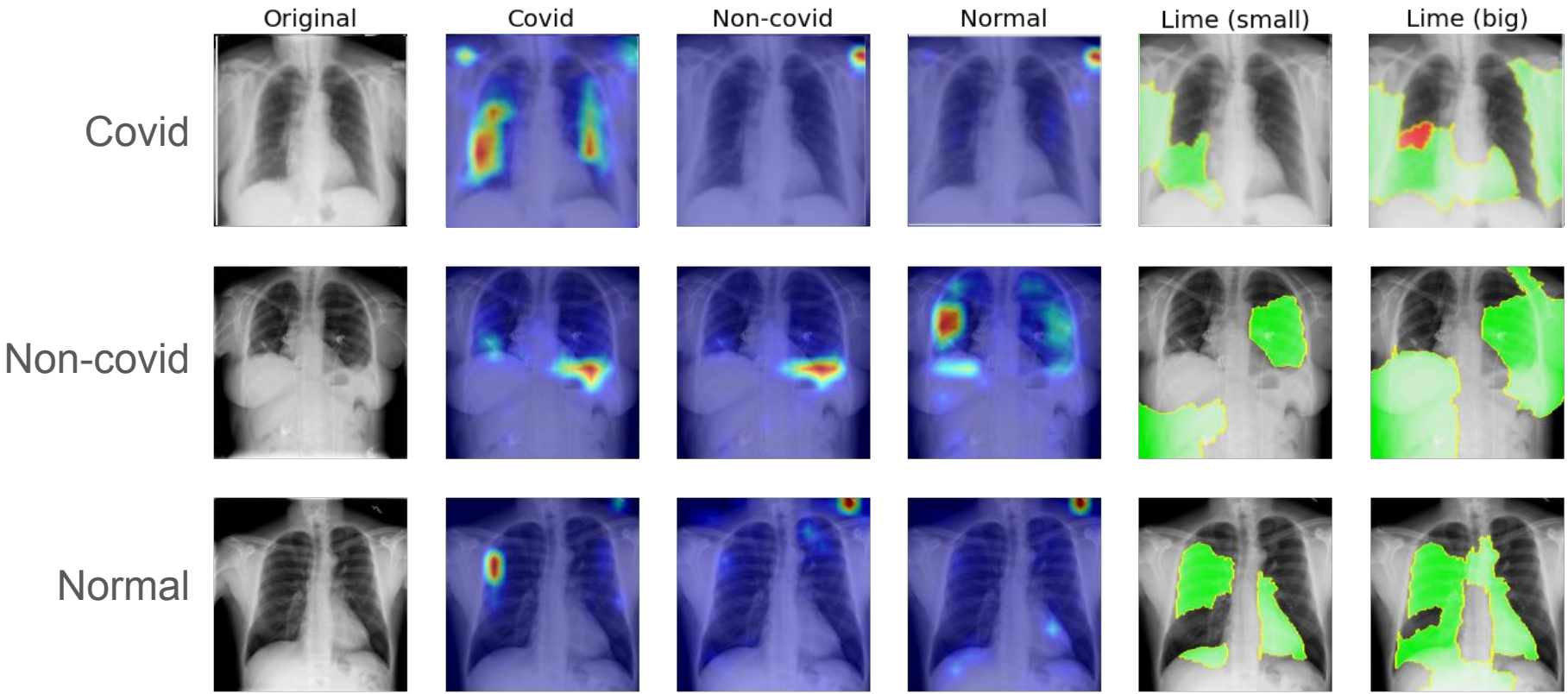
Kajetan Husiatyński, Piotr Komorowski, Szymon Antoniak

2. LRP for saliency maps

1. Train model



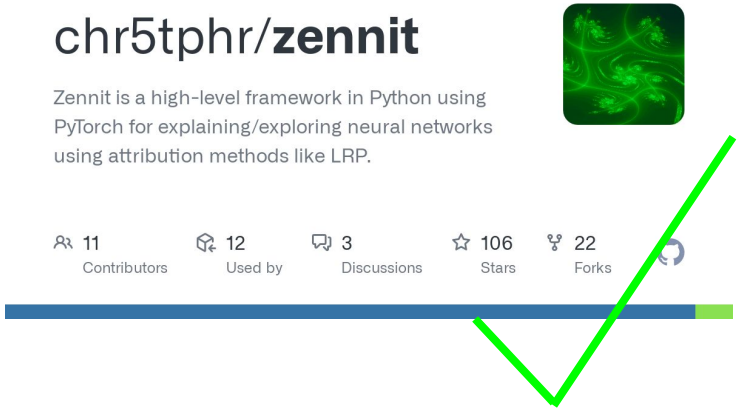
Visual Results



Quantitative results

Explanation type	Faithfulness Correlation (avg)	Average Sensitivity (avg)
baseline (LIME)	0.35	0.45
TransformerLRP	0.47	0.34

Quantus - bugfix




```
x.register_hook(self.save_inp_grad)
```

```
def predict(self, x: np.ndarray, grad: bool = True, **kwargs) -> np.array:
    """
    Predict on the given input.

    Parameters
    ----------
    x: np.ndarray
        A given input that the wrapped model predicts on.
    grad: boolean
        Indicates if gradient-calculation is disabled or not.
    kwargs: optional
        Keyword arguments.
```

Key takeaways

- We trained a Vision Transformer for Medical Imaging (Covid vs other lung infection vs healthy)
- A novel relevance propagation explanation (Chefer et al., 2020) designed specifically for Transformers was used to visualise features learned by the model
- The explanations were validated using the Quantus package (required significant engineering effort due to the explanation not being native to Quantus)
- The explanation method produces explanations that are higher quality compared to baselines, both visually and metric-wise