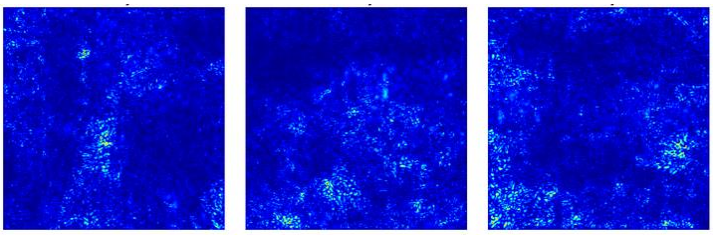
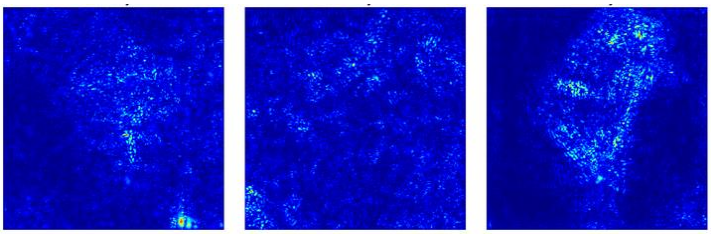
Report

Part 1:

1. The saliency maps to me are not that useful, since they do not clearly show what is happening behind the dots, and they do not show clear outlines of objects or something of the like. The Grad-Cam does show better insights, because it creates more of an overlay. I think the explanation for library 2 is not correct, since it 'identified' a lamp, a chair and the edge of the table and categorized it as a harp, but in the explanation you cannot clearly see what part of the picture influences this decision most strongly.
2. I think the explanation could still be useful, since you can see what parts of the image influenced the decision, although the decision itself is not correct.
3. 
4. I have done this for mushr1, mushr2 and para1. Saliency map 2 is not that useful since it looks like it does not use the pixels that actually show the object. the other two do show the pixels on top of the object. the grad cam clearly shows that 1 and 3 are interpreted correctly, while 2 has identified as a gold ball, mostly because it has chosen one of the mushrooms (which is round and white) as a predictor for that class.

5. Gradcam seems like the most intuitive one, because it shows general areas of interest. Smoothgrad is nice to have, because it shows individual pixel importance, but because of this it is less useful, because when looking as a human at an image you interpret a set of pixels as an image and not the individual pixels. Saliency maps are in my opinion even less useful because of this same reason, but the saliency shows the picture importance in even less detail.

Part 2:

1. The idea of counterfactual explanations are useful and easy to understand, because they basically state what minimal change needs to be made to change the prediction. In practice, this could mean for example for a cancer patient that when a counterfactual predicts that smoking less would change the prediction greatly, that gives an actionable step to take. Intuitively, counterfactuals are pretty clear as well (to the general public and to me).
2. I can see that counterfactuals can lead to more trust. LIME can provide insights into what features are important for a given prediction, but it can also be too abstract for non-experts. It needs some interpretation to makes sense as well, which people might

not have. Counterfactuals are more intuitive and therefore better for this. The only downside I am seeing is that a drastic change in behaviour because of a counterfactual estimation for a different decision in life to change the predicted outcome could lead to someone seeing it as too drastic of a change, making them trust it less, even though they know that it is a meaningful (positive) change.