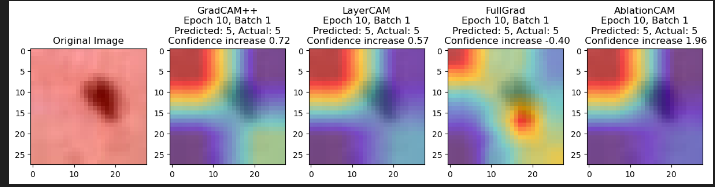
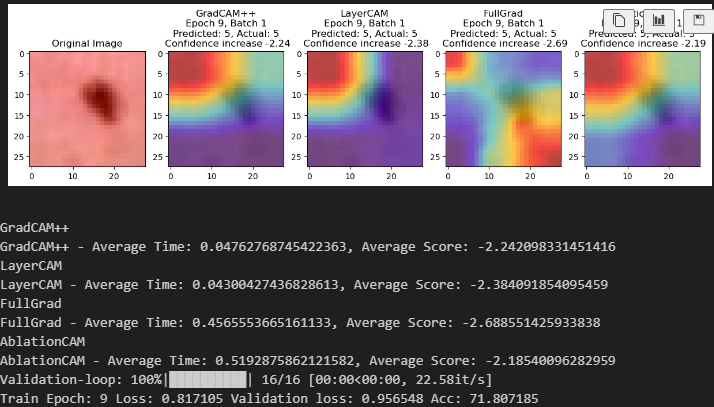
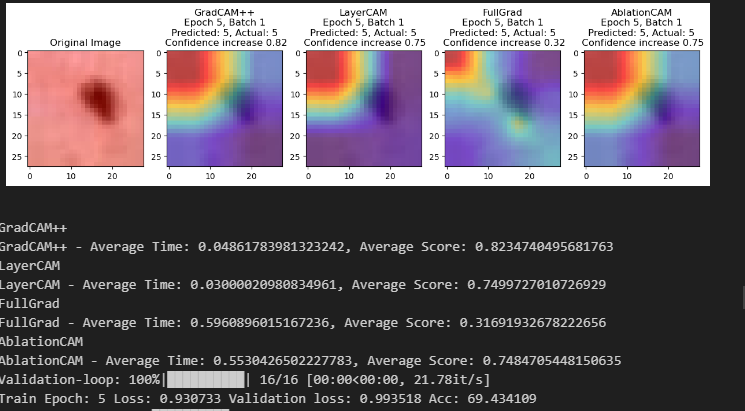
3-5 methods

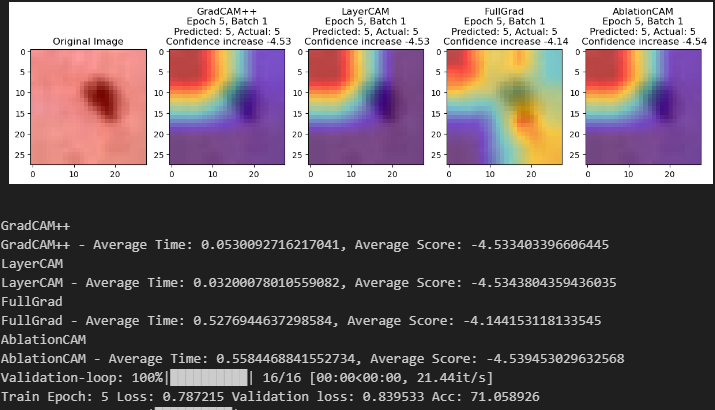
DermaMNIST

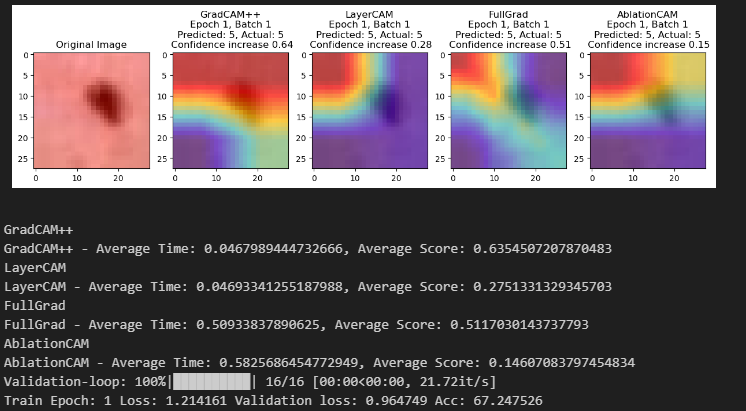
Epoch 10

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | GradCAM++ | LayerCAM | FullGrad | AblationCAM |
| ROADCombined | 0.72 | 0.57 | -0.4 | 1.96 |
| Avg. Time | 0.05 | 0.04 | 0.46 | 0.53 |
| Human evaluation |  |  |  |  |
| Convident increase | -2.24 | -2.38 | -2.69 | -219 |
| Model ACC: 71. 57 |  |  |  |  |
|  |  |  |  |  |

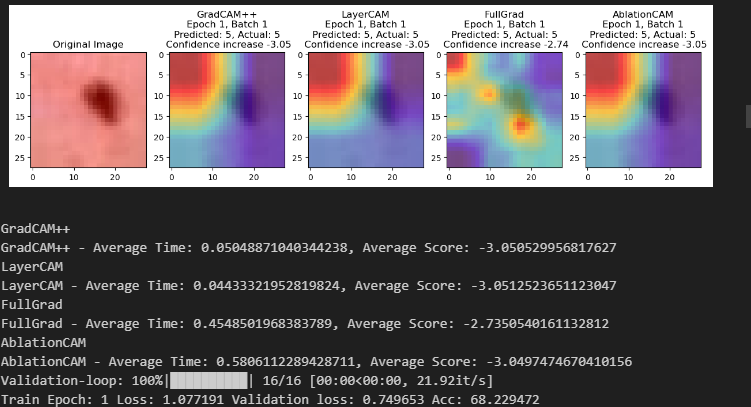
ROADCombined:   


ConfidenceChange after Softmax





Change after softmax:

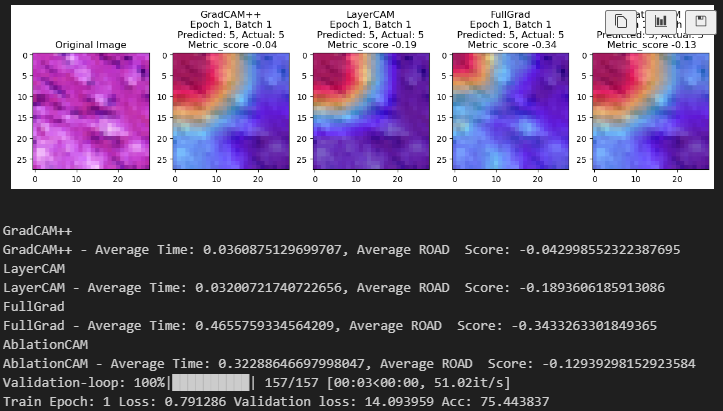


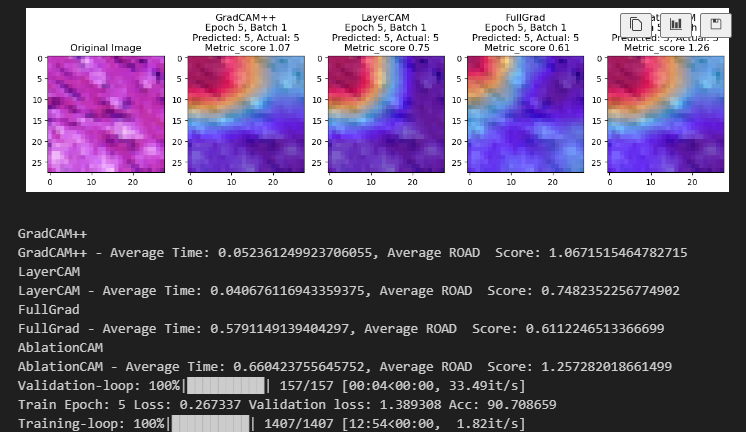
PathMNIST

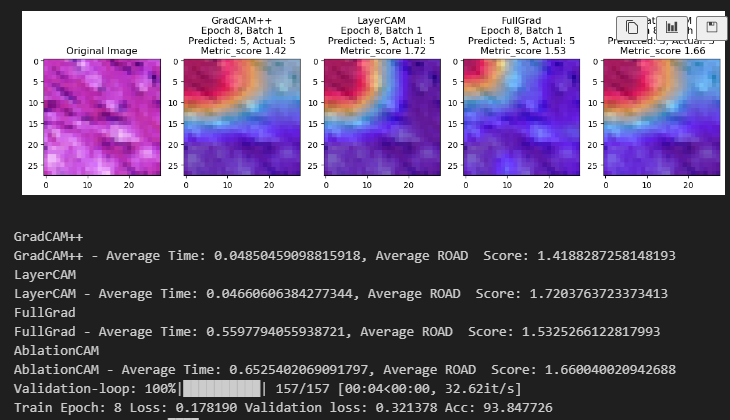
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | GradCAM | GradCAM++ | LayerCAM | FullGrad |  |
| ROADCombined |  |  |  |  |  |
| Avg. Time |  |  |  |  |  |
| Human evaluation |  |  |  |  |  |
| Convident increase |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |

Image from validation run in last epoch:

Epoch 1:



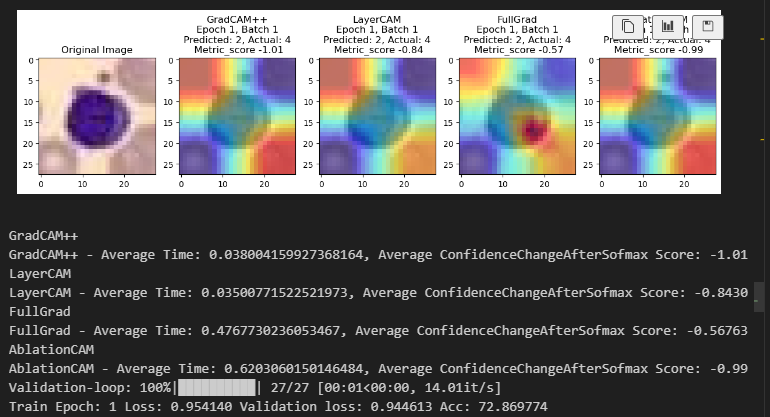


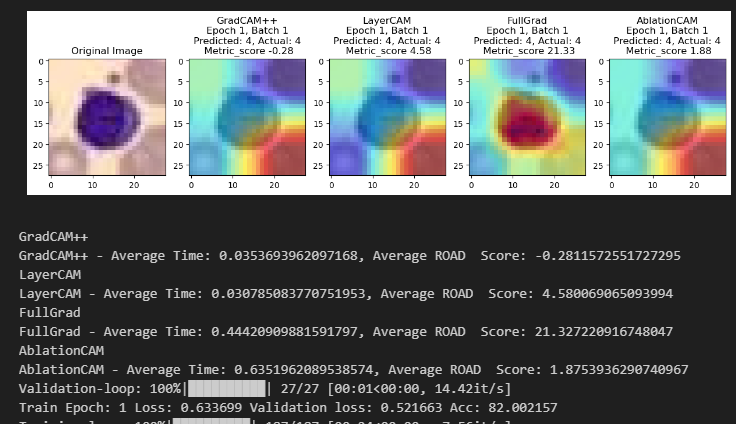


BloodMNIST

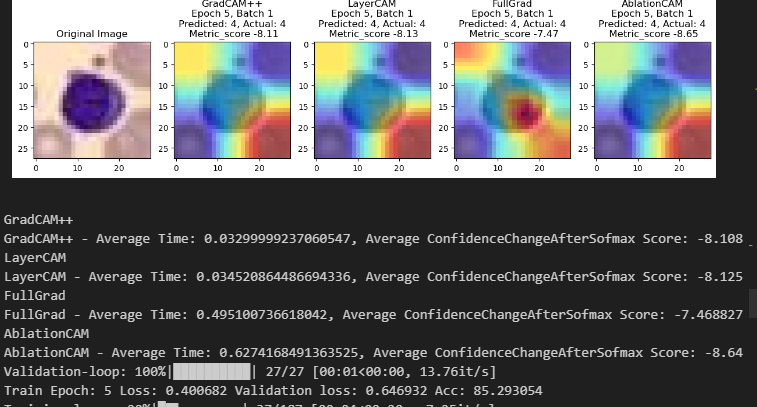
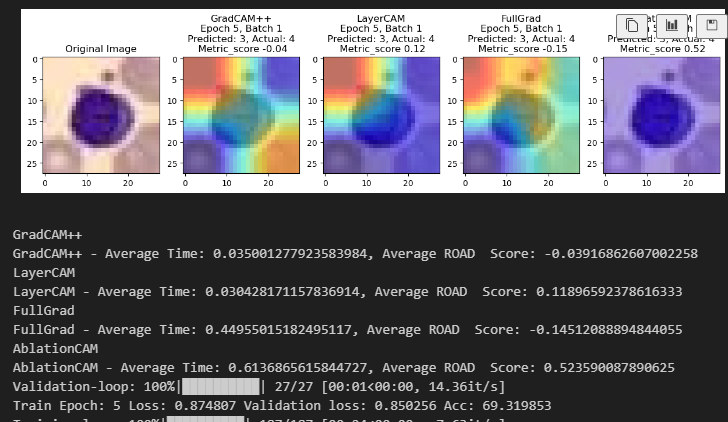
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | GradCAM++ | LayerCAM | FullGrad | AblationCAM |
| ROADCombined | -0.04 | 0.12 | -0.14 | 0.52 |
| Avg. Time | 0.04 | 0.03 | 0.45 | 0.61 |
| Human evaluation |  |  |  |  |
| Convident increase | 1.66 | 1.61 | 2.52 | 1.59 |
|  |  |  |  |  |

Epoch 1

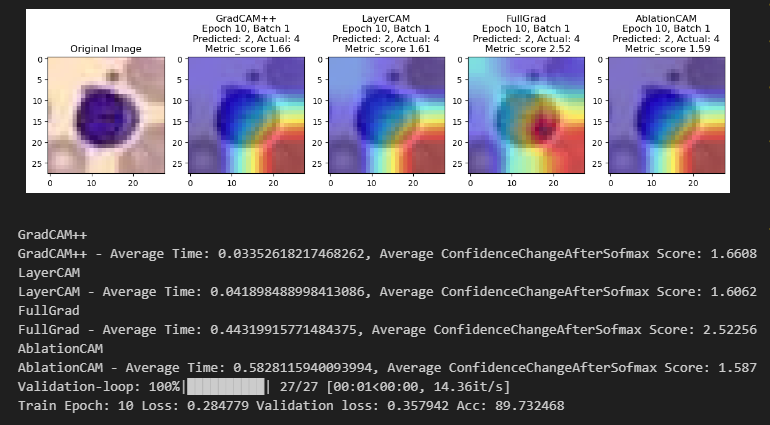


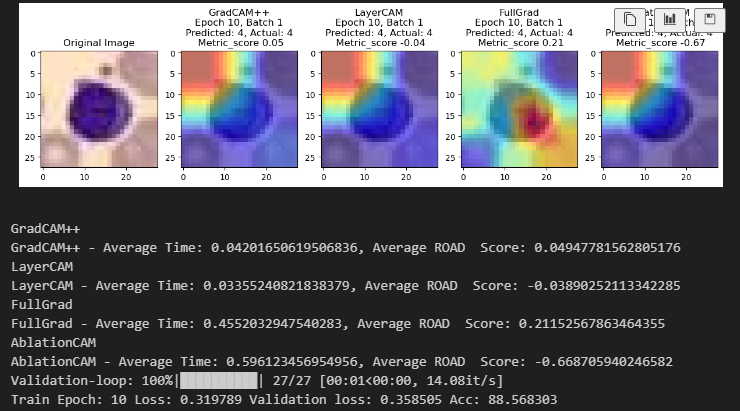


Epoch 5:

Epoch 10





Record images

Record metrics (class-specific accuracy)

**How to assess the performance of the different CAM-methods:**

There are different options to do this. The first would be to compare ground truth annotations for what regions of the image are important for each class to the CAM output. However, as there are no ground truths provided, Intersection over Union (IoU) or Jaccard Score are not metrices we could use.

Another possibility is qualitative assessment: By adding a function which prints the CAM-Method output overlayed on the image to predict on users (or experts) can now rate the quality of the CAM outputs in terms of interpretability and alignment with human understanding.

This we, even though we have no specific knowledge about blood images, did.

The third metrics we came up with is Computational Efficiency: By recording the time taken by each CAM method to process an image we evaluated how efficient each method is.

Then we used the framwork provided and evaluated the ROADCombined(percentiles=[20, 40, 60, 80]) metrics. This a combined evaluation from multiple perspectives, considering how well the CAM highlights pixels across different activation thresholds.

A higher score would typically indicate that the CAM is effectively highlighting the pixels most relevant to the model's prediction, while a lower score might indicate that the CAM is less effective or highlighting irrelevant areas. That is why we decided to give the CAM with the highest ROADCombined score the most points.

As second quantitative method we want to measure the change in the confidence, after softmax, that's why we also used ClassifierOutputSoftmaxTarget. And then logged the confidence increase/decrease.

Which specific method outperforms the rest (performance, similarity to how you would explain the image)

**Which dataset & models**

We used three different datasets DermaMNIST; BloodMNIST and PathMNIST. With the DermaMNIST set as smallest set for initial training and longer evaluation. The BloodMNIST as comparison dataset to see if the same CAM methods provide similar results on a different dataset. And the PathMNIST as largest and more classes to evaluate if an increase in classes has an impact on the CAM methods. We used a standard resnet50 implementation, but also played around with a CNN architecture we coded ourselves.

Table with ranking of explanation methods

Figures demonstrating ROIs

Compare prediction from start of training to end of training

* Substantial difference between the features that are selected for explanation?
  + For PathMNIST the features selected for explanation stayed quite constant throught the epochs
  + For DermaMNIST LayerCAM and AblationCAM used similar features selected for explanation throught the training while FullGrad and GradCAM++ changed quite a lot which features they selected for explanation
  + For BloddMNIST the FullGrad method selected as only method different aspects for explanation. All the other, especially GradCAM++ and LayerCAM agreed on the features selected for explanation.
* At what point can one stop training because features look good enough
  + For PathMNIST after epoch 7 this can be related to the way larger training set than in the other datasets
  + For both BloodMNIST and DermaMNIST in between the CAM’s output different features selected for explanation, therefore the point at which the training should stop would be when the validation score does not improve anymore.