

# Figure 4.4: Counterfactual Validation Examples for Grad-CAM Attribution

Original Pair

Attribution Map  
(Grad-CAM)

High-Attribution  
Masked

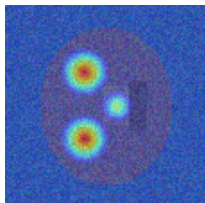
Low-Attribution  
Masked

Verdict

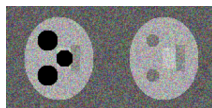
$s_{\text{orig}} = 0.78$



Grad-CAM



$s = 0.40$   
 $\Delta s = -0.38$



Large score drop

$s = 0.73$   
 $\Delta s = -0.05$



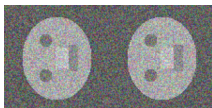
Small score drop



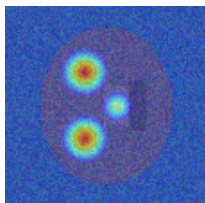
NOT FALSIFIED

Attribution correctly identified important features

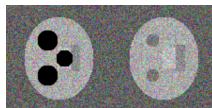
$s_{\text{orig}} = 0.62$



Grad-CAM

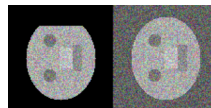


$s = 0.50$   
 $\Delta s = -0.12$



Moderate drop

$s = 0.51$   
 $\Delta s = -0.11$



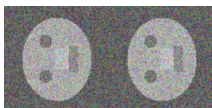
Moderate drop



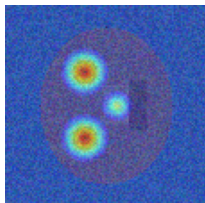
FALSIFIED

Attribution failed - no differential prediction

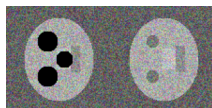
$s_{\text{orig}} = 0.45$



Grad-CAM

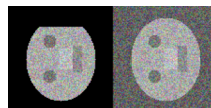


$s = 0.17$   
 $\Delta s = -0.28$



Large score drop

$s = 0.41$   
 $\Delta s = -0.04$



Small score drop



NOT FALSIFIED

Crosses verification threshold ( $0.45 \rightarrow 0.17$ )