Deep Learning Lab-6

immediate

November 7, 2024

Abstract

In this Lab, we integrated the transfer-learned ResNet model, re-trained on the MexCulture dataset (from Lab 01), with multiple explanation techniques previously developed in earlier labs. Specifically, we used this ResNet model as the backbone classifier to apply the four explanation methodsâGRADCAM, FEM, RISE, and LIME (developed in Labs 4 and 5). These methods were employed to generate saliency maps, offering insight into the classifier's focus areas on the images.

The generated saliency maps were then evaluated using two categories of metrics. The first category, which we had explored in Lab 02, includes the Pearson Correlation Coefficient (PCC) and Structural Similarity Index (SIM), providing a quantitative comparison to a ground truth. The second category introduces two new metrics, Insertion and Deletion, to assess the robustness of the explanations without requiring ground truth.

1 Explanation Models

Explanation models clarify the decision-making process of complex models, making their predictions more understandable. They provide insights in a way humans can easily interpret, helping to make blackbox models more transparent.

1.1 Whitebox mMdels

They assist users in understanding how and why a model makes specific predictions, which is especially important in applications where interpretability is critical

1.1.1 GRADCAM (Gradient Class Activation Mapping)

Grad-CAM is a post-hoc visualization technique used to explain the class-discriminative regions of an image that contribute to a modelâs prediction. It leverages the structure of Convolutional Neural Networks (CNNs) to create a heat map highlighting the important pixels that influence the prediction for a particular class. Grad-CAM works by analyzing the feature maps from the last convolutional layer, as deeper layers in CNNs capture high-level features. It computes the gradient of the class score with respect to these feature maps, which helps identify the regions of the image most relevant to the prediction.

1.1.2 FEM (Feature-based Explanation Method)

FEM is similar to Grad-CAM in that it also utilizes the high-level features extracted by deeper convolutional layers in a CNN. However, FEM focuses on identifying strong and rare features that influence the final prediction. It assumes that these features follow a Gaussian distribution and employs K-sigma filtering to detect them from the last convolutional layer. By isolating these influential features, FEM provides a clearer understanding of how the model arrives at its decisions.

1.2 Blackbox Models

Opaque models, often referred to as blackbox models, make predictions in a way that is not easily interpretable by humans. Their internal logic, rules, and processes are complex and opaque, making them difficult to understand.

1.2.1 LIME

LIME is a feature attribution method that determines the importance of each feature in an input sample for a modelâs prediction. It works by perturbing the input sample and analyzing the corresponding changes in predictions to identify important features.

1.2.2 RISE

RISE (Randomized Input Sampling for Explanation) generates a saliency map to highlight the important features for a modelâs prediction. It creates numerous low-resolution binary masks, upscales them, and evaluates the modelâs prediction for each masked input. Unlike other methods, RISE assumes unmasked features are crucial for the prediction if the model's output remains strong after masking.

2 Dataset

We have been provided with the images `img` and `test_saliency_img`, and in addition, we will use the MexCulture142 dataset, which consists of 284 images of Mexican monuments, along with corresponding gaze fixation points and ground truth Gaze Fixation Density Maps (GFDM). The dataset is divided into training and validation sets, with images labeled according to their filename prefixes.

3 Workflow of the Code

The code workflow begins by loading test images and, if available, their corresponding Ground-Truth Gaze Fixation Density Maps (GDFMs). Each image is preprocessed and passed through a pretrained model (e.g., ResNet) to generate saliency maps using specified methods such as GRADCAM, FEM, RISE, or LIME. The generated saliency maps are normalized, resized, and visualized with heatmaps and overlays on the original images.

3.1 Evaluation with Ground Truth

When ground truth is available, it calculates the **Pearson Correlation Coefficient (PCC)** and **Structural Similarity Index (SSIM)** to compare the generated saliency maps with the GDFMs. This comparison evaluates how closely the saliency maps align with the ground-truth data.

Evaluation without Ground Truth

It evaluates the saliency maps using the **Insertion** and **Deletion** algorithms, without ground-truth data:

- **Insertion:** This algorithm progressively adds pixels to a blurred version of the image, based on their importance as indicated by the saliency map. The prediction score is recalculated at each step, and the area under the curve (AUC) is computed to quantify the performance.
- **Deletion:** This algorithm progressively removes pixels from the image, starting with the most important ones. The model's prediction score is tracked, and the AUC is used to evaluate how the removal of these pixels affects the model's confidence.

Table 1: Parameters and their values for LIME and RISE methods

Parameter	**Value**
positive_only	True
negative_only	False
num_superpixels	15
hide_rest	True
low_res_mask_size	8
mask_number	200
threshold	0.4
size	(224, 224)
top_labels	1
hide_color	[0, 0, 0]
num_lime_features	100000
num_samples	2000
rand_index	0

4 Evaluation with GRADCAM

The **GRADCAM** explanation model was tested using the entire dataset with a transfer-learned **ResNet** model across all three image classes. For evaluation **with ground truth**, metrics such as **PCC** and **SSIM** were used, comparing the generated saliency maps to the Ground-truth Gaze Fixation Density Maps (GDFMs) to assess their alignment with human attention. Additionally, the entire dataset was analyzed **without ground truth** using **Insertion** and **Deletion** metrics to evaluate how the model's prediction confidence changed as important pixels were added or removed. This thorough approach provided comprehensive insights into both the external validity of the saliency maps compared to human gaze data and their effectiveness in reflecting the model's internal feature importance.

Note: Insertion and Deletion AUC are calculated without ground truth for all evaluations.

Table 2: Summary of GradCAM Metrics

Metric	Mean	Variance
PCC	0.570	0.150
SSIM	0.655	0.051
Insertion AUC	0.584	0.398
Deletion AUC	0.359	0.323

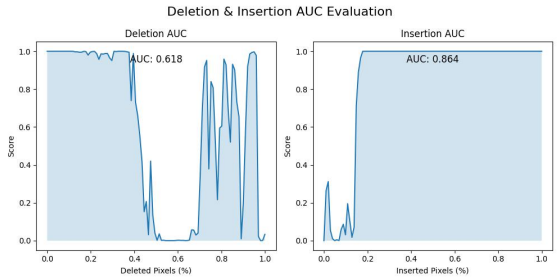


Fig. 1: Insertion and Deletion AUC.

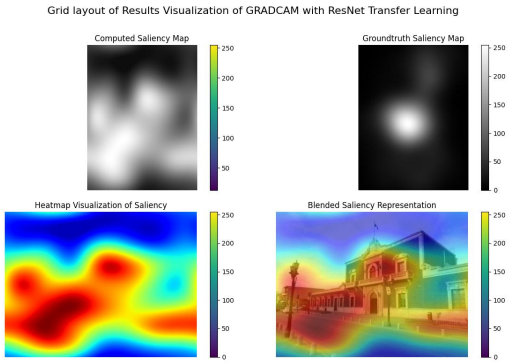


Fig. 2: GradCAM Visualization Result with ResNet.

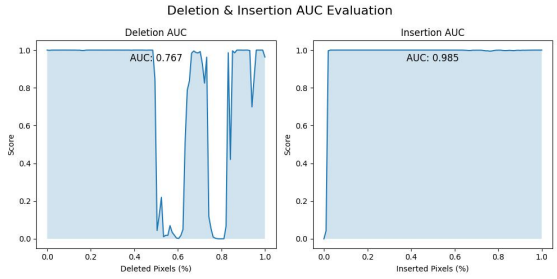


Fig. 3: Insertion and Deletion AUC.

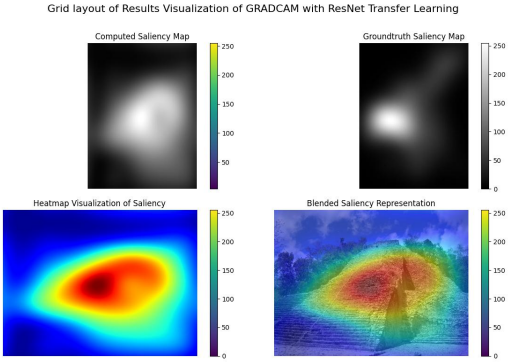


Fig. 4: GradCAM Visualization Result with ResNet.

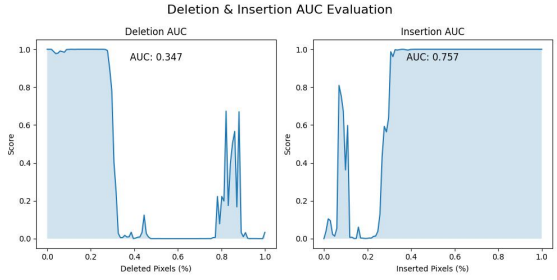


Fig. 5: Insertion and Deletion AUC

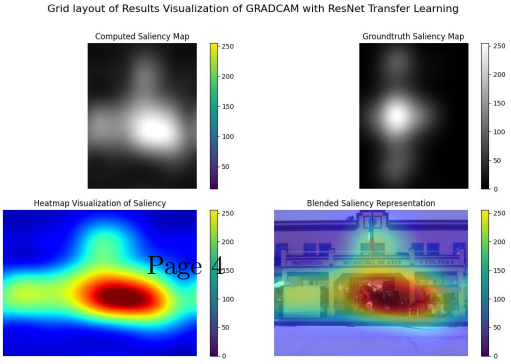


Fig. 6: GradCAM Visualization Result with ResNet.

4.1 GradCAM Results

The GradCAM-generated saliency maps show well-defined, focused regions of importance, particularly in Figures 2 and 4, which align closely with the ground-truth GDFMs. However, Figure 6 displays broader, more dispersed highlights, indicating occasional variability in precision.

- **PCC**: The mean PCC of **0.570** indicates a moderate positive correlation between GradCAM saliency maps and the ground truth, showing that relevant features are captured but with room for improvement. A variance of **0.150** suggests consistent performance across images.
- **SSIM**: The mean SSIM of **0.655** demonstrates reasonable structural similarity between the GradCAM maps and the ground truth, with a low variance of **0.051**, indicating stability in visual feature representation.
- **Insertion AUC**: The mean of **0.584** implies GradCAM effectively highlights relevant features that positively impact model confidence, though the high variance of **0.398** points to fluctuating performance.
- **Deletion AUC**: A mean of **0.359** reflects less effective identification of irrelevant features, with a high variance of **0.323** indicating inconsistencies in suppressing unimportant areas.

4.2 GradCam Figures Result Explanation

- **Fig.1**: The deletion AUC of 0.618 indicates moderate capability in suppressing irrelevant features, while the insertion AUC of 0.864 shows strong recognition of key features, despite some inconsistency.
- **Fig.3**: The deletion AUC of 0.767 reflects improved suppression of non-important areas, and the insertion AUC of 0.985 highlights excellent feature recognition, showing reliable performance overall.
- **Fig.5**: The deletion AUC of 0.347 reveals difficulties in ignoring irrelevant features, whereas the insertion AUC of 0.757 indicates effective and suggests the model can still identify relevant regions but variable identification of key regions.

5 FEM

The FEM explanation model was tested on the entire dataset using a transfer-learned ResNet model across all three image classes. For ground truth evaluation, PCC and SSIM metrics were applied to compare the generated saliency maps to the Ground-truth Gaze Fixation Density Maps (GDFMs) to measure alignment with human attention. The dataset was also analyzed without ground truth using Insertion and Deletion metrics to assess changes in model confidence when important pixels were added or removed.

Table 3: Summary of GradCAM Metrics

Metric	Mean	Variance
PCC	0.570	0.150
SSIM	0.655	0.051
Insertion AUC	0.584	0.398
Deletion AUC	0.359	0.323

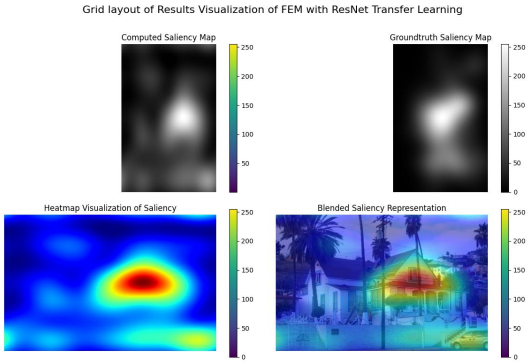


Fig. 7: Insertion and Deletion AUC.

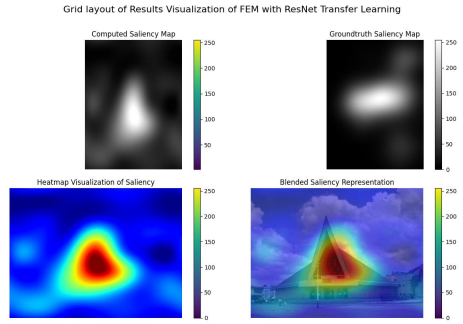


Fig. 8: FEM Visualization Result with ResNet.

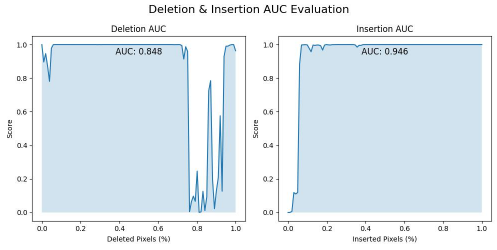


Fig. 9: Insertion and Deletion AUC.

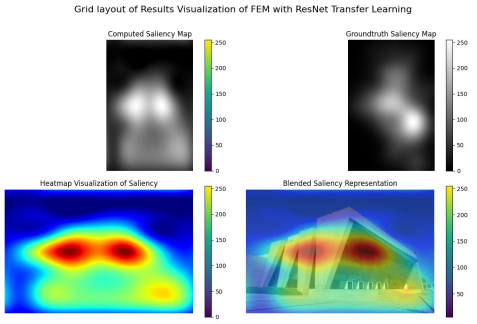


Fig. 10: FEM Visualization Result with ResNet.

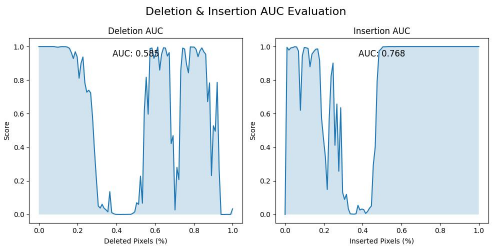


Fig. 11: Insertion and Deletion AUC.

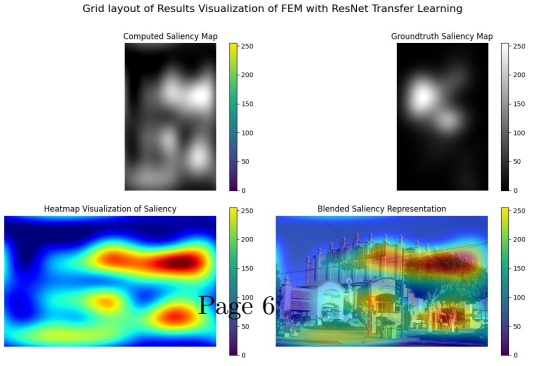
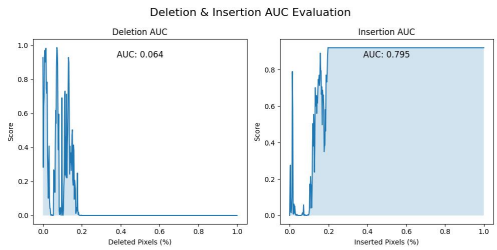
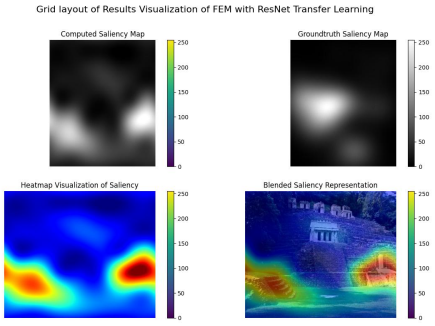


Fig. 12: FEM Visualization Result with ResNet.





5.1 FEM Results

- **PCC and SSIM**: The PCC mean (0.433) and SSIM mean (0.595) suggest the model has a moderate ability to align its saliency maps with human attention, but it is not very strong. The variances are relatively high (0.244 for PCC and 0.081 for SSIM), indicating that this alignment is inconsistent across images. Overall, this isn't ideal but not completely poor.
- **Insertion AUC**: With a mean of 0.565, the model shows a fair ability to identify and highlight important features. However, the high variance (0.395) indicates that this ability varies significantly between different images, suggesting that the model sometimes struggles to highlight key areas effectively.
- **Deletion AUC**: The mean of 0.376 is quite low, showing that the model has difficulty suppressing irrelevant features when important pixels are removed. The high variance (0.341) suggests that its performance is inconsistent, pointing to unreliable suppression of non-important content.

Note: For more detailed explanations of the figures, please refer to page 12, section 7.2.

6 RISE

The RISE explanation model was evaluated on a total of 9 images, comprising 3 images from each of the three image classes, using a transfer-learned ResNet model.It also has ground truth evaluation where PCC and SSIM metrics were applied to compare the generated saliency maps to the Ground-truth Gaze Fixation Density Maps (GDFMs) to measure alignment with human attention. The dataset was also analyzed without ground truth using Insertion and Deletion metrics to assess changes in model confidence when important pixels were added or removed.

Table 4: RISE Performance Metrics on the Entire Dataset

Metric	**Mean**	**Variance**
PCC	0.027	0.205
SSIM	0.559	0.073
Insertion AUC	0.388	0.155
Deletion AUC	0.339	0.139

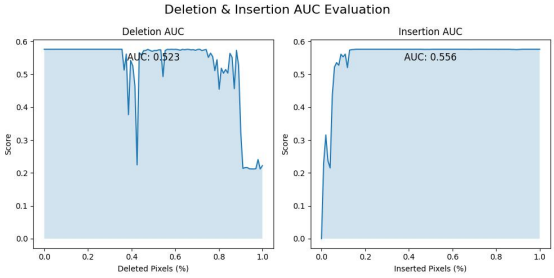


Fig. 13: Insertion and Deletion AUC.

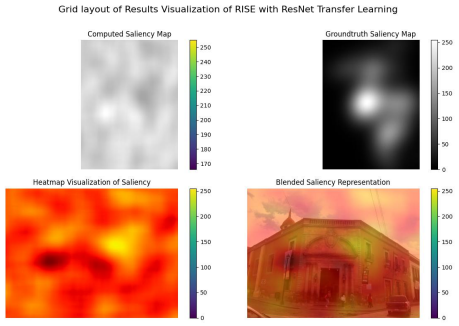


Fig. 14: RISE Visualization Result with ResNet.

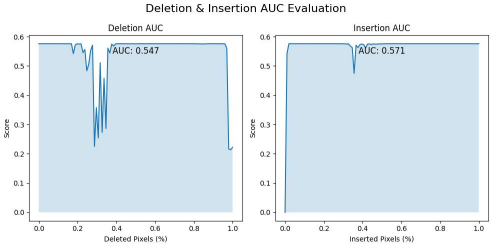


Fig. 15: Insertion and Deletion AUC.

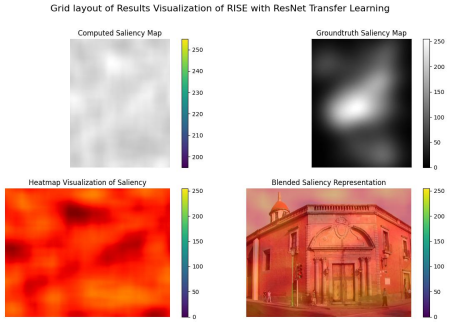


Fig. 16: RISE Visualization Result with ResNet.

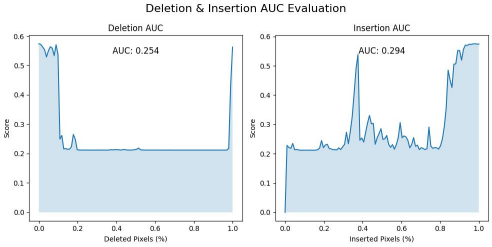


Fig. 17: Insertion and Deletion AUC.

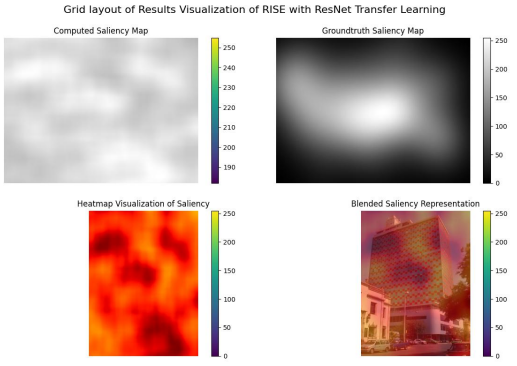


Fig. 18: RISE Visualization Result with ResNet.

6.1 RISE Results

- **PCC (mean: 0.027, variance: 0.205)**: The very low mean of the PCC suggests a weak correlation between the generated saliency maps and the ground truth, indicating poor alignment with human attention. The higher variance implies considerable inconsistency across different samples.
- **SSIM (mean: 0.559, variance: 0.073)**: The SSIM mean indicates a moderate similarity between the saliency maps and ground truth. The relatively low variance points to consistent performance across different samples, though the similarity level is not particularly strong.
- **Insertion AUC (mean: 0.388, variance: 0.155)**: The insertion AUC is relatively low, implying that the model struggles to correctly highlight important features. The variance is moderate, suggesting that there is some inconsistency in the model's ability to identify relevant regions.
- **Deletion AUC (mean: 0.339, variance: 0.139)**: The deletion AUC is also low, showing that the model has difficulty in correctly suppressing unimportant features. The moderate variance indicates a somewhat inconsistent ability to filter out non-essential areas.

6.2 RISE Figures Result Explanation

The RISE-generated saliency maps highlight regions in the images but often appear diffused and not sharply defined, especially in Figures 14 and 18. The ground-truth saliency maps (GDFMs) are more focused, indicating that RISE may struggle to capture relevant features with precision. Figure 16 shows better alignment with more concentrated highlights, while Figures 14 and 18 reveal broader, generalized regions lacking specificity seen in human attention. However, Insertion and Deletion AUC (Figures 13, 15, 17) Insertion AUC values indicate that adding important pixels only partially restores model confidence. Deletion AUC values are consistently low, especially in Figure 17, implying that removing identified pixels does not significantly lower confidence, suggesting less relevant feature identification.

7 LIME

The LIME model was evaluated on 9 images (3 per class) using a transfer-learned ResNet model. Ground truth evaluation was performed with PCC and SSIM metrics to compare saliency maps to GDFMs, measuring alignment with human attention. Insertion and Deletion metrics were also used without ground truth to assess changes in model confidence when important pixels were added or removed.

Table 5: LIME Model Performance Metrics

Metric	Mean	Variance
PCC	0.113	0.226
SSIM	0.573	0.085
Insertion AUC	0.452	0.166
Deletion AUC	0.276	0.112

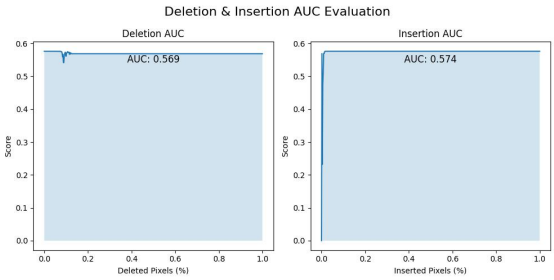


Fig. 19: Insertion and Deletion AUC.

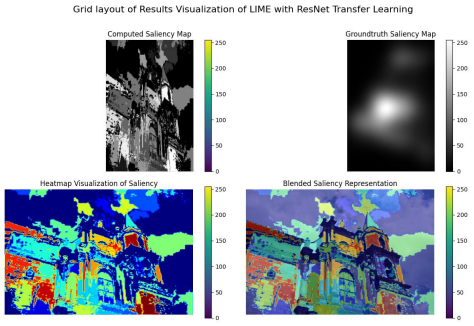


Fig. 20: LIME Visualization Result with ResNet.

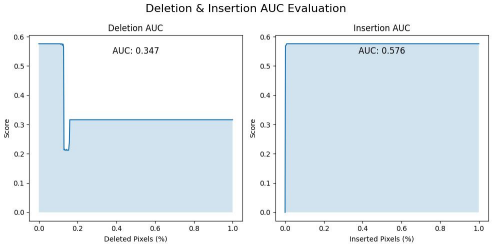


Fig. 21: Insertion and Deletion AUC.

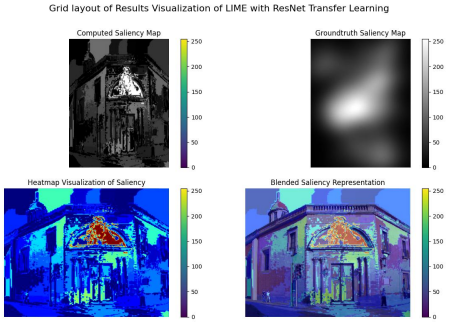


Fig. 22: LIME Visualization Result with ResNet.

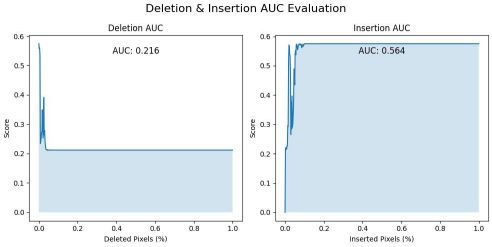


Fig. 23: Insertion and Deletion AUC.

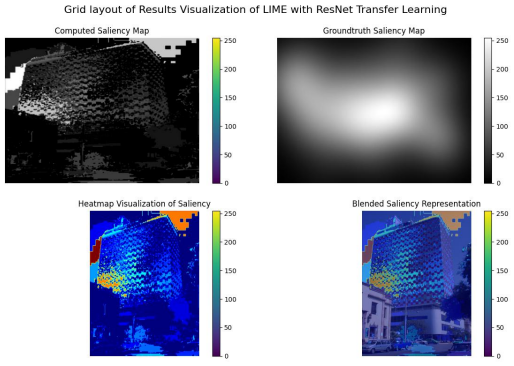


Fig. 24: LIME Visualization Result with ResNet.

- **PCC (mean: 0.113, variance: 0.226)**: The mean value suggests a very weak correlation between the generated saliency maps and the ground truth, implying that the model's explanations do not align well with human attention. The relatively high variance indicates inconsistency across different samples.
- **SSIM (mean: 0.573, variance: 0.085)**: The Structural Similarity Index (SSIM) shows a moderate similarity between the saliency maps and the ground truth, suggesting that some visual consistency is maintained. The low variance indicates that the model's performance is relatively stable across instances.
- **Insertion AUC (mean: 0.452, variance: 0.166)**: This metric shows that the model's ability to identify and emphasize important features is moderate. The moderate variance suggests some variation in performance, with certain instances where the model performs better or worse.
- **Deletion AUC (mean: 0.276, variance: 0.112)**: The lower mean for deletion AUC implies that the model struggles with correctly ignoring or removing irrelevant features. The low variance here indicates that this performance is consistent, though poor, across samples.

7.1 LIME Figures Result Explanation

- **Saliency Maps (Figures 20, 22, 24):**
  The LIME-generated saliency maps display highlighted regions, but they do not consistently match significant features that one would expect based on visual intuition. The ground-truth saliency maps (GDFMs) are more concentrated and focused, indicating where human attention likely falls. In contrast, LIME's results show fragmented and less focused regions, suggesting inconsistency in identifying the most relevant areas influencing the model's prediction.
- **Insertion and Deletion AUC (Figures 19, 21, 23):**
  The Deletion AUC values are low, which implies that removing the pixels LIME identifies as important does not significantly decrease the model's confidence. This suggests that LIME may not be pinpointing truly impactful features. The Insertion AUC values are somewhat better but still not strong, indicating that adding back the "important" pixels identified by LIME only moderately increases the model's confidence, which is less than ideal for explaining model behavior.

7.2 FEM Figures Result Explanation

The FEM-generated saliency maps highlight regions in the images that are more focused and defined, as seen in Figures 8 and 12, where the maps align well with ground-truth saliency maps (GDFMs) and show concentrated highlights capturing key features. However, Figure 10 shows broader, less precise areas, indicating occasional overestimation. The Insertion and Deletion AUCs (Figures 7, 9, 11) reflect this, with high Insertion AUC values suggesting that adding identified pixels restores model confidence significantly, while moderate to high Deletion AUC values indicate that removing these pixels lowers confidence, affirming the importance of the features identified by FEM. These results imply that FEM effectively highlights critical areas impacting the modelâs predictions, though there is some variability in precision.

8 Conclusion

The lab session aimed to integrate code blocks, implement new saliency map evaluation techniques, and evaluate four explanation methods (GradCAM, FEM, RISE, and LIME) using a transfer-learned ResNet model on the MexCulture dataset. This included analyzing Insertion and Deletion metrics.

- **GradCAM**: This method demonstrated strong performance across most metrics, including high Insertion AUC and moderate PCC and SSIM values. It was particularly effective in highlighting important features, though it showed some variability in handling irrelevant features. Overall, GradCAM provided the most balanced and robust results among the methods tested.
- **FEM**: The FEM method showed a moderate correlation with human attention as indicated by PCC and SSIM values. It performed well in distinguishing key features, reflected in its higher Insertion AUC, but struggled with maintaining consistency in ignoring unimportant features (lower Deletion AUC). Despite these limitations, FEM demonstrated reliable visual explanations.
- **RISE**: The RISE model presented challenges, with lower mean PCC and Insertion AUC values, suggesting weaker alignment with ground truth and less effective feature identification. The moderate SSIM indicates some visual consistency, but overall, this method performed less effectively compared to GradCAM and FEM.
- **LIME**: LIME had the weakest performance among the evaluated methods. Its low PCC and moderate SSIM indicate limited correlation and visual similarity with ground truth, while the Insertion and Deletion AUC values highlighted its struggles in both identifying relevant features and ignoring irrelevant content. LIME's higher variance suggests inconsistent performance across different samples.

9 References

- Ayyar, M. P., Benois-Pineau, J., & Zemmari, A. (2021). White Box Methods for Explanations of Convolutional Neural Networks in Image Classification Tasks. *arXiv preprint arXiv:2104.02548*.
- Fuad, K. A. A., Martin, P. E., Giot, R., Bourqui, R., Benois-Pineau, J., & Zemmari, A. (2020, November). Features understanding in 3D CNNs for actions recognition in video. In *2020 Tenth International Conference on Image Processing Theory, Tools and Applications (IPTA)* (pp. 1-6). IEEE.
- Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 618-626).