Mini-Report: LIME-Based Analysis of Smile vs. Not Smile Predictions

Jader Silva

Introduction

This section documents the interpretability results for both classes (smile and not smile) in a toy neural network trained on synthetic faces. Using LIME explanations, we analyzed a balanced set of correctly classified examples: smile (indices 110, 66, 32, 41, 116) and not smile (indices 54, 69, 86, 93, 77), each predicted with probability 1.00. The objective was threefold: (1) to identify recurring visual patterns emphasized by the model, (2) to compare how explanations differ between positive and negative classes, and (3) to detect potential spurious signals such as border fixation or background noise. In addition, we examined the single misclassified instance in the test set (idx 132) to contrast correct reasoning with failure cases and highlight the role of data quality in interpretability.

Method

Samples: A balanced selection of correctly classified examples from both classes (smile and not smile), each predicted with probability 1.00, plus one misclassified case for contrast.

Tool: lime_image.LimeImageExplainer with 800 perturbations per image to generate local explanations.

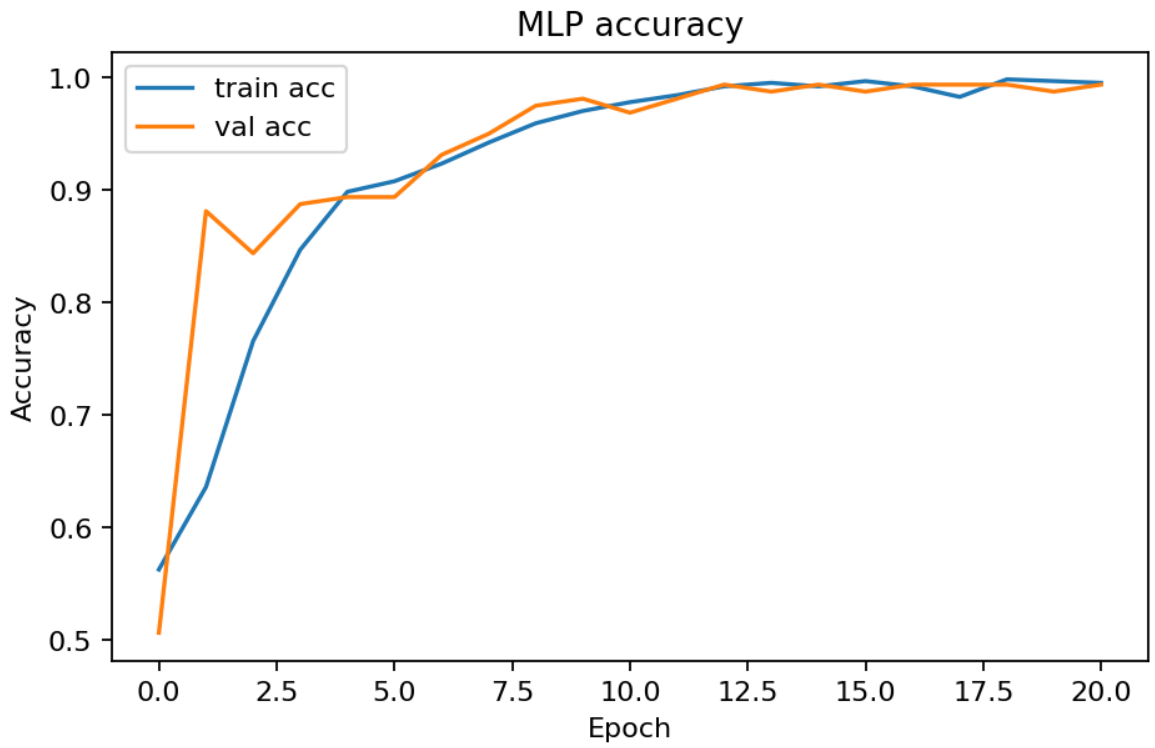
Analysis: Each LIME output was manually inspected for four key aspects: (1) mouth highlights, (2) eye/upper-face involvement, (3) border fixation, and (4) noise or background artifacts.

Criteria: Consistency of focus across samples, plausibility of highlighted regions as human-like reasoning, and differences between correct vs. incorrect predictions.

Visualizations

I generated several visualizations to evaluate model performance and interpretability:

Figure 1.

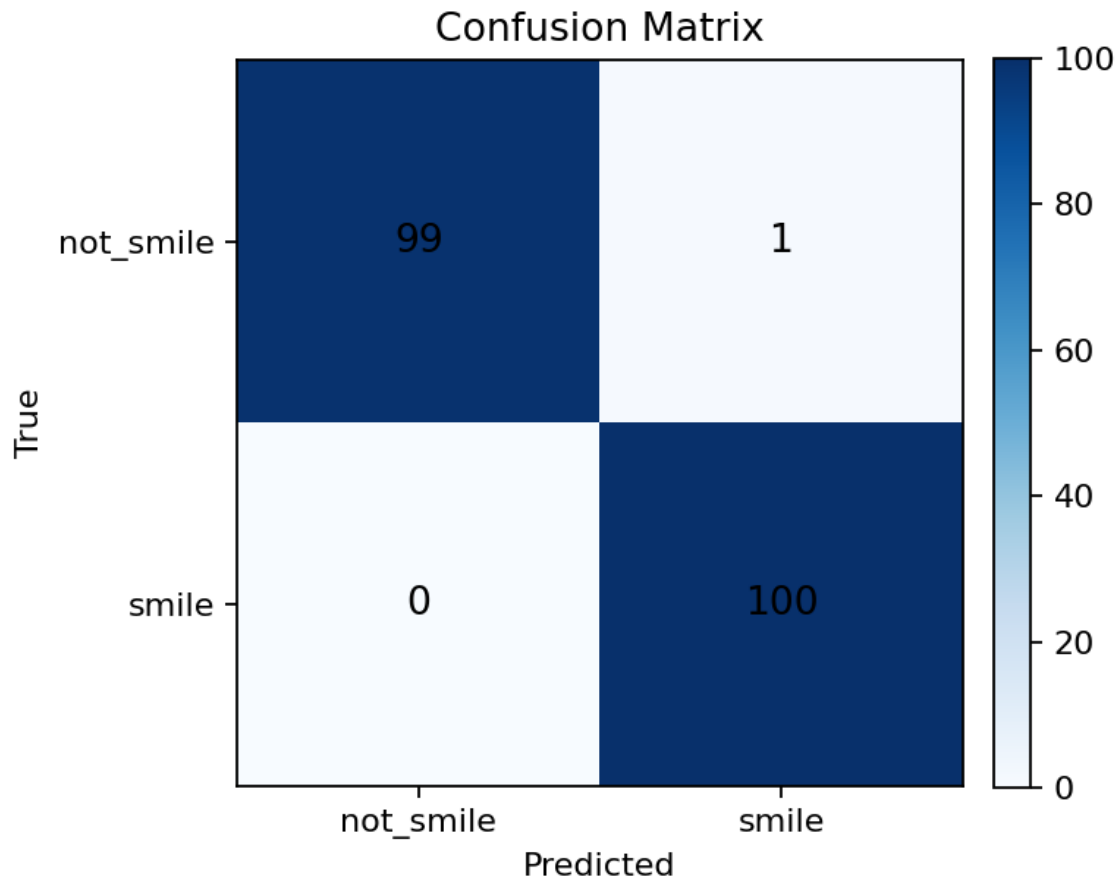


Early phase (0–2.5 epochs):
Validation accuracy (orange) rises very quickly because the model captures the easiest patterns right away.
Training accuracy (blue) increases more slowly as the network adjusts its weights.
This creates the initial "isolation" of the orange curve.

Around epoch 2.5:
The two curves meet and begin to oscillate closely together, crossing multiple times after epoch 10.
This interweaving shows there is no strong overfitting: training and validation remain nearly aligned.

After epoch 10:
Both curves stabilize near 0.97–0.98 accuracy.
They "dance" around each other, alternating which one is slightly higher.
This reinforces that the model has converged successfully and is not collapsing.

Figure 2.



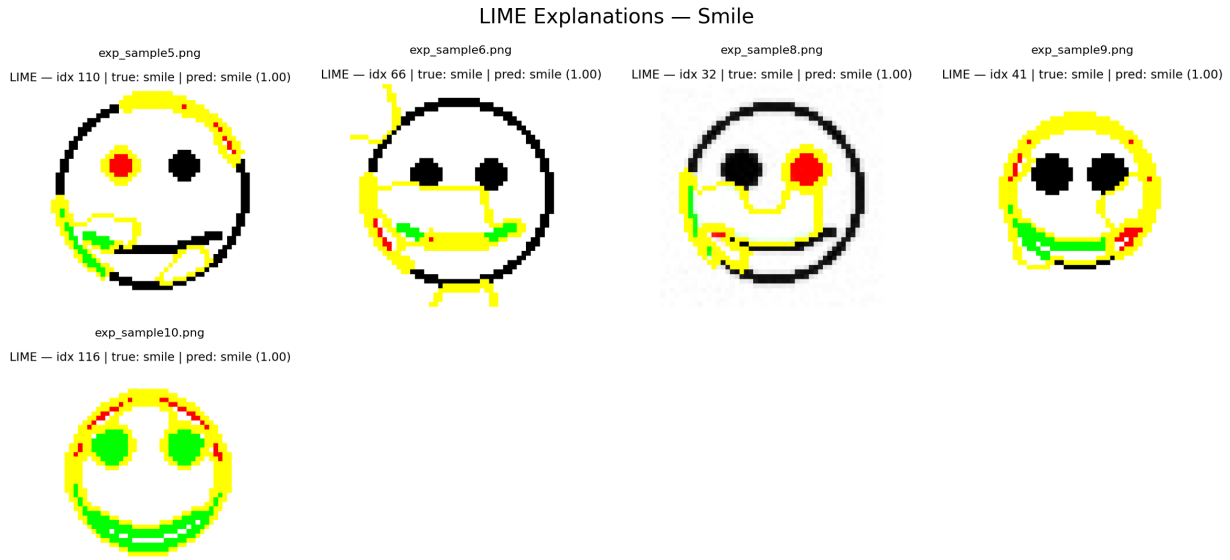
Confusion matrix – Out of 200 test samples (100 per class), the model correctly classified 199. The single error occurred when a 'not smile' icon (idx 132) was predicted as 'smile'. This imbalance (no misclassified smiles, only one misclassified not-smile) suggests the model is slightly more confident distinguishing smiles than rejecting false positives.

Figure 3. Smile Grid (LIME) – highlights mouth curves as primary signals, with occasional border fixation.

Figure 4. Not Smile Grid (LIME) – highlights upper-face and eyes more often than mouth.

Figure 5. Error Case (LIME) – reveals spurious cues in the only misclassified instance.

Findings (Figure 3. Smile Grid)



1. Sample 110 (idx 110)

- **Highlights**: Mouth curve (green/yellow), left jawline, red/yellow on eye and top of head.
- **Interpretation**: Correctly identifies smile curve but mixes it with secondary focus on eyes and head border.

2. Sample 66 (idx 66)

- **Highlights**: Mouth (yellow with green corners), vertical strip along left side of face, yellow bottom, top scribble noise.
- **Interpretation**: Smile detected, but strong attention to contour and scattered noise show sensitivity to irrelevant pixels.

3. Sample 32 (idx 32)

- **Highlights**: Mouth (yellow, red corner), vertical strip on left face side, eye fully red with yellow halo.

- **Interpretation**: Smile curve is correct feature, but eye is heavily overweighted — spurious correlation risk.

4. Sample 41 (idx 41)

- **Highlights**: Mouth (green with red corner), full yellow circle around head, scattered red patches.
- **Interpretation**: Smile detected, but explanation dominated by border fixation — classification right, reasoning questionable.

5. Sample 116 (idx 116)

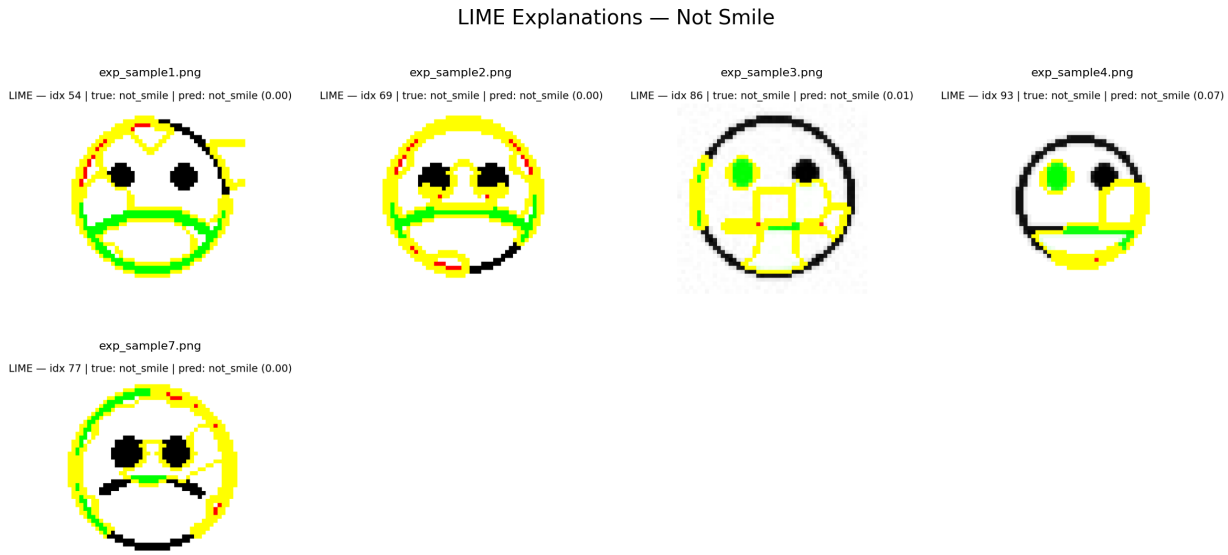
- **Highlights**: Mouth fully green, both eyes green/yellow, full yellow border, red arc at top.
- **Interpretation**: Strong focus on mouth (good), but overemphasis on eyes and head outline suggests redundant and noisy attention.

Comparative Summary

- **Consistent signal:** All five samples highlight the mouth curve as the strongest feature, aligning with human intuition.
- **Secondary features:** Eyes and head borders are often highlighted, sometimes disproportionately (e.g., sample 32, 116).
- **Border fixation:** Two cases (41, 116) show strong yellow outlines, suggesting reliance on global shape rather than smile-specific features.
- **Noise artifacts:** Top scribbles or random red patches appear in samples 66 and 41.

While the model demonstrates reliable detection of smiles through the mouth curve, LIME reveals a tendency to include irrelevant borders and eye regions, highlighting the importance of explainability to detect hidden biases.

Findings (Figure 4. Not Smile Grid)



1. Sample 54 (idx 54)

- **Highlights:** Lower half of face strongly green, surrounded by thick yellow; red accent on left upper head (viewer's perspective).
- **Interpretation:** Model captures the frown curve correctly, but extends attention broadly across the lower face. Head-border emphasis introduces possible noise.

2. Sample 69 (idx 69)

- **Highlights:** Mouth fully green, surrounded by yellow; large thick yellow covering ~80% of head, red patches on both upper sides and near chin; yellow also touches both eyes.
- **Interpretation:** Correct detection of mouth as frown signal, but excessive yellow on the head border and eyes indicates overgeneralization. Risk of spurious correlation with head outline rather than facial expression.

3. Sample 86 (idx 86)

- **Highlights:** Left side thick yellow with green patch, left eye fully green with yellow halo; random yellow scribbles across face, including mouth area; small red dots near mouth and center.

- **Interpretation:** Highly noisy explanation. While left eye and part of the mouth receive focus, scattered highlights suggest instability in explanation. Model still outputs correct label, but reasoning appears weak.

4. Sample 93 (idx 93)

- **Highlights:** Mouth fully highlighted, especially right side upper curve in green; yellow extends upward from mouth to right eye; left eye fully green with yellow halo.
- **Interpretation:** Strong evidence from mouth (good), but heavy involvement of both eyes indicates reliance on mixed cues. Correct classification, but reasoning may not be strictly expression-based.

5. Sample 77 (idx 77)

- **Highlights:** Left arc of face in green with yellow outline; opposite side fully yellow with scattered red dots; mouth minimally marked (only center with small green/yellow); yellow around eyes and between them.
- **Interpretation:** Weak mouth focus, with majority of attention on borders and eyes. Model arrives at correct classification, but explanation dominated by irrelevant regions.

Comparative Summary

- **Consistent signal:** Mouth curve is present in most explanations, but often diluted compared to Smile Grid.
- **Secondary features:** Eyes receive strong attention across multiple samples (54, 86, 93, 77), often more dominant than the mouth.
- **Border fixation:** Similar to Smile Grid, yellow frequently outlines the head (69, 77), raising concern about reliance on global shape rather than expression.
- **Noise artifacts:** Random scribbles and scattered red/yellow dots appear in 86 and 77, indicating instability in explanation.

Overall, while the model successfully distinguishes “not smile,” LIME explanations reveal overemphasis on eyes and head borders rather than consistently prioritizing the frown curve. This suggests the classifier is accurate but may not always rely on robust, human-intuitive features.

Findings (Figure 5. Error Case & Data Quality)

- **Observation**: The single misclassified instance (idx 132) presents a flat, straight mouth line that does not clearly correspond to either smile or not_smile.
- **Interpretation**: Although labeled as not_smile, the geometry is ambiguous (closer to a neutral expression). The classifier predicted smile with 0.75 probability, and LIME highlights confirm attention to the horizontal mouth line. This suggests the model responded consistently, but the data itself introduced uncertainty.
- **Takeaway**: This misclassification is not a true reasoning failure of the model, but an artifact of **dataset quality and labeling ambiguity**.
- **Next Steps**: To address this, future iterations could (a) enforce stricter curvature rules during synthetic data generation, or (b) introduce an explicit neutral class. Both options would reduce confusion and improve label fidelity.

Results

The evaluation confirmed that the model reached stable convergence, as shown by the accuracy curve and confusion matrix. Training and validation accuracies intertwined closely after epoch 2.5 and stabilized around 0.97–0.98 after epoch 10, confirming that the model generalized consistently without signs of overfitting. These metrics provide a reliable baseline for interpreting the LIME explanations. Building on this, the analysis of LIME grids revealed several key patterns.

Consistent visual cues:

- For smile, the model consistently emphasized the mouth curve as the strongest and most human-intuitive feature.
- For not smile, the mouth was also present but often diluted by competing signals from eyes and head borders.

2. **Differences between positive and negative classes**:

- Smile explanations generally aligned with human reasoning, yet secondary highlights (eyes, head outline) appeared frequently, sometimes dominating the explanation (e.g., samples 32, 41, 116).
- Not smile explanations leaned more heavily on borders and eye regions, with the frown curve receiving less exclusive attention (e.g., samples 54, 69, 77). This suggests that the model distinguishes “not smile” less by the absence of a curve and more by global head or eye-based cues.

3. **Spurious signals**:

- Border fixation was observed in both classes, especially in samples 41, 69, and 77, where thick yellow rings captured attention regardless of facial expression.
- Noise artifacts such as scattered red dots and scribble-like patches (samples 66, 86, 77) revealed instability, showing that the model can latch onto irrelevant pixels.

4. **Error case (idx 132)**:

- This misclassified instance corresponds to the single error in the confusion matrix. The 'not smile' icon presented a flat, ambiguous mouth line closer to neutral than to either smile or not smile. LIME confirmed that the classifier relied on this horizontal mouth bar, and the ambiguous labeling introduced uncertainty. This shows that the error stems more from dataset limitations than from flawed reasoning by the model.

Summary

Overall, the results confirm that the model can reliably separate smiles from non-smiles, but with systematic biases. While smiles are detected through the intuitive mouth curve, non-smiles often

depend on less robust cues such as eye shading and head borders. The error case underscores the importance of data quality, as ambiguous inputs can mislead even a well-trained classifier. These findings align with the Mini-Report's objectives, providing evidence of recurring patterns, contrasting reasoning strategies between classes, and exposing spurious signals that affect interpretability.

Ethical Reflections

The interpretability patterns uncovered in this toy study raise broader concerns when extrapolated to real-world domains such as facial recognition and justice systems. In facial recognition, reliance on spurious cues (e.g., borders or irrelevant artifacts) mirrors risks of demographic bias, where models may anchor decisions on superficial features rather than genuine expressions or identities. Similarly, in justice applications, such as forensic analysis or courtroom evidence, explanations that emphasize noise instead of core signals highlight the danger of overconfidence in algorithmic outcomes. The error case (idx 132), where ambiguity in data labeling led to misclassification, illustrates how fragile evidence becomes if model reasoning is not aligned with human-intuitive features. These findings underscore the need for transparent interpretability tools to expose hidden biases, ensuring that algorithmic decisions affecting individuals' rights are both explainable and accountable.

At the same time, the ambiguous misclassification in this study reminds us that data quality is inseparable from justice. In real-world settings, poor lighting in surveillance footage or degraded image quality in forensic evidence can create similar ambiguities. If a model is allowed to "decide" on such uncertain inputs without transparent interpretability, its output may be misused as an objective fact, when in reality it rests on unstable ground. This reinforces the ethical imperative: in both facial recognition and justice systems, models must not only achieve high accuracy but also demonstrate how and why their decisions are made, so that human oversight remains central in safeguarding fairness and accountability.

Limitations

Despite providing useful insights, this study has several limitations. First, the dataset consists of synthetic faces with simplified geometry. While this makes the model's reasoning easier to visualize, it also reduces the complexity of real-world variability (e.g., lighting, pose, occlusion, and demographic diversity). As a result, findings may not fully generalize to real images of human faces.

Second, the explanations obtained through LIME are themselves approximations. Perturbation-based methods are known to produce unstable results, sometimes highlighting irrelevant regions. This means that while the patterns we observed (e.g., border fixation, eye involvement) are meaningful, they should be interpreted cautiously and validated with complementary methods.

Third, the analysis focused on a small number of representative samples (five per class plus one misclassified case). This allowed for detailed qualitative interpretation, but limits statistical robustness. Broader sampling would be required to ensure that the highlighted tendencies are systematic rather than anecdotal.

Finally, the single error case (idx 132) revealed that dataset quality directly affects interpretability. Ambiguous or poorly rendered samples may distort both model predictions and explanations, making it difficult to distinguish between genuine reasoning errors and data artifacts.

Next Steps

Building on these findings, several directions can strengthen both the technical and ethical dimensions of the project.

1. **Dataset refinement**: Future iterations should improve data generation by enforcing stricter curvature thresholds for smiles and non-smiles, thereby reducing ambiguous or neutral samples. An additional "neutral" class could also be introduced, reflecting a broader spectrum of facial expressions.

2. **Broader sampling and validation**: Expanding the number of analyzed samples per class would allow for stronger conclusions about systematic patterns. Quantitative metrics (e.g., frequency of border fixation across explanations) could complement the qualitative findings presented here.

3. **Alternative interpretability methods**: While LIME provides valuable local explanations, exploring complementary approaches such as SHAP, Integrated Gradients, or Grad-CAM could help validate or challenge the observed patterns, reducing the risk of over-interpreting perturbation noise.

4. **Application to real-world images**: Moving beyond synthetic data, the same pipeline could be applied to real facial expression datasets. This would test whether the spurious signals identified here (border fixation, eye dominance, noise artifacts) also emerge in more complex and socially relevant contexts.

5. **Link to critical domains**: Finally, the interpretability framework developed in this toy study can be extended to ethically sensitive applications such as facial recognition and justice systems. Examining how models behave in those contexts — and whether they rely on robust or spurious cues — will be essential for ensuring fairness, transparency, and accountability.