Emotion Classification in Dogs Using ResNet and LLM

Abstract

Understanding and detecting emotional states in domestic dogs holds substantial significance across industries such as pet care, veterinary medicine, and animal welfare. However, dog emotion recognition remains challenging due to subjective interpretations and variability in visual cues. This project aimed to develop an AI-powered system that classifies dog emotions from images into three categories: Happy, Sad, and Angry. We implemented and compared three deep learning models: a customized Convolutional Neural Network (CNN), MobileNetV2, and a fine-tuned ResNet18 architecture. The models were trained on an augmented dataset designed to address class imbalance and visual variability. Among them, ResNet18 achieved the highest performance with a test accuracy of 84.0% and a macro-averaged ROC-AUC score of 0.955.

To enhance interpretability, we integrated Grad-CAM visualizations and natural language explanations generated by a large language model (LLM), resulting in a transparent and user-friendly web-based application. This combination of visual and textual interpretability ensures that the system is not only accurate but also accessible to non-technical users, fostering trust and usability. The integration of Grad-CAM highlights the features most relevant to the model's predictions, while the LLM provides human-readable explanations, making the system transparent and interpretable.

This system offers significant business value by addressing key pain points in industries focused on animal welfare and pet care. In pet care and veterinary services, the system enables early detection of happiness, angriness or sadness, allowing for timely intervention that can reduce risks, enhance safety, and improve care quality. For smart pet monitoring solutions, integrating emotion recognition provides real-time insights for pet owners, creating differentiated products with emotionally intelligent features. In animal shelters and training facilities, the system supports staff in identifying and managing behavioral issues, leading to better welfare outcomes and improved adoption rates.

Looking forward, expanding the range of emotions detected, incorporating multimodal data sources such as video and audio, and fine-tuning the LLM on domain-specific datasets will further enhance the system's accuracy, interpretability, and market potential. By combining cutting-edge AI with a focus on transparency, this project lays the foundation for practical and trustworthy dog emotion recognition systems with broad applications in pet care, veterinary medicine, and animal welfare.

The full implementation and additional resources can be accessed via our GitHub repository: https://github.com/Alphteow/DoggyEmotion.

1. Introduction

Understanding animal emotions, particularly those of domestic dogs, due to their close interactions with humans, has significant implications across industries such as pet care, veterinary medicine, dog training, and animal welfare. Dogs are highly expressive animals, yet their emotional states are often misinterpreted or entirely overlooked, particularly in settings where human-animal interaction is brief, high-volume, or emotionally charged. In environments like doggy daycares, shelters, kennels, and veterinary clinics, early recognition of negative emotions such as sadness and angriness can reduce the risk of injury, illness, and behavioral issues. Timely intervention not only improves animal welfare but also enhances operational safety and service quality.

Despite the clear value, emotion detection in dogs remains largely subjective, as emotional expressions can vary greatly between individual animals. Our project aims to address this challenge by developing an AI-powered tool that provides a consistent, image-based second opinion on canine emotions, using visual data such as facial expressions and body posture. Rather than replacing human judgment, the system offers real-time insights that can assist staff and pet owners in making more informed decisions to support a dog's well-being.

To solve this problem, we explore deep learning-based image classification models that interpret dog emotions from images. The goal is to design a system that is not only accurate, but also interpretable and accessible to non-technical users. Our final solution integrates model predictions, visual explanations (via Grad-CAM), and natural language rationales (via a large language model) into a user-friendly interface.

2. Related Work

Facial emotion recognition has been a central task in computer vision, with applications ranging human computer interaction to behavioral analysis. While most research has traditionally focused on humans, recent work has extended these techniques to animal emotion recognition, particularly in animals like dogs and cats. In this domain, deep learning models have been used to identify subtle facial cues associated with emotions such as happiness, sadness, anger, and fear.

To address challenges of data scarcity and visual variability in model training, Shorten and Khoshogoftaar (2019) provided a comprehensive survey of augmentation strategies in deep learning and highlighted their effectiveness across a range of vision-based tasks. Our work follows this practice by applying augmentation to expand the training set, ensuring better class representation and more stable training dynamics

In terms of model architecture, CNNs have been widely adopted. Mao and Liu (2023) proposed a CNN-based framework optimized with a whale optimization algorithm for dog facial expression recognition, achieving superior performance over traditional architectures. Similarly, Wu et al. (2023) applied deep learning techniques, including ResNet, to classify emotions in both dogs and cats, demonstrating the applicability of general purpose classification networks in pet emotion detection.

Recent advancements have introduced the use of large language models (LLMs) to generate natural language explanations for deep learning predictions. For example, Lu et al. (2024) proposed an Emotion-Action Interpreter powered by an LLM (EAI-LLM), which provides detailed, human-readable justifications for emotion classification based on 3D body movement data. This illustrates a growing trend toward combining visual recognition with language generation to enhance interpretability and user trust in AI systems.

3. Dataset and Preprocessing

3.1 Dataset

The original dataset consists of labelled images of dogs, each grouped by emotion categories such as Happy, Angry, and Sad. In total the dataset contains 4730 images across folders, with a significantly imbalanced distribution. The ‘Happy’ and ‘Angry; class contains 1865 images, while ‘Sad’ includes only 1000 images.

To assess data quality, several checks were performed on the raw dataset. No exact duplicates were found using perceptual hashing. An exposure analysis revealed that little to no images were underexposed or overexposed, which may have affected visual clarity. Additionally, the images displayed a wide range of resolutions and aspect ratios, highlighting the need for consistent resizing and padding during the preprocessing.

Each image varies in lighting, framing, and expression, making the dataset diverse but also noisy. A sample image from each of the three primary emotion classes (Happy, Sad, Angry) is shown in Figure 1 for visual reference. Overall, while the dataset provides a foundational starting point for emotion classification, it requires augmentation and normalization to be suitable for robust deep learning workflows.

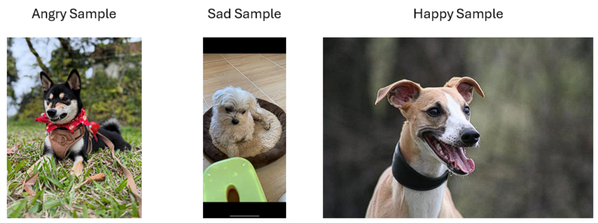


Figure 1. Sample image from the dataset after augmenting and processing for each emotion class

3.2 Preprocessing and Augmentation

To prepare the raw dog emotion dataset for deep learning, all images were first standardized to a uniform resolution of 224 x 224 pixels using a combination of resizing and padding. Since the original images varied in aspect ratio, each image was resized while maintaining its original proportions and then centered on a black canvas to match the target shape. This ensured consistent input dimensions without distorting the dog’s facial features.

Following this preprocessing step, each image was augmented three times to enhance dataset diversity and improve model generalization. Two types of label-preserving transformations were applied, fixed angles rotations and flipping operations. Fixed angles of 90°, 180°, or 270° were randomly applied to simulate the potential variations in camera orientation during real-word image capture. Moreover, each image was also flipped using one of three modes of horizontal, vertical, or transpose. Changes in viewpoint, such as mirroring or head tilting, do not alter the emotional cues of a dog’s expression.

These augmentations aim to replicate natural variations in dog posture and camera perspective while preserving the integrity of each emotional label. As a result of our data collection, the Happy class increased to 6,254 images, the Angry class to 6,018 images, and the Sad class to 3,236 images. Because sad dog images were more challenging to obtain, we deliberately kept the smaller dataset for the Sad class to reflect real-world conditions when training the model.

By artificially increasing intra-class variability, the model is encouraged to learn more expression specific visual patterns rather than overfitting to specific pose or lighting. This strategy is especially important given the relatively small and imbalanced nature of the raw dataset, as it helps mitigate overfitting and improves the models performance on unseen data.

3.3 Exploratory Data Analysis(Post-Augmentation)

EDA was conducted on the post-augmentation dataset to evaluate how the preprocessing and augmentation pipeline affected data quality, class balance, and feature consistency across emotion classes. Post augmentation analysis is particularly important because the model no longer trains on the raw dataset, it learns from transformed inputs instead. Therefore, verifying the distribution and visual features of the processed dataset ensures that the augmentations introduce meaningful diversity without compromising label integrity or inducing biases.

Following augmentation, the dataset was significantly expanded as shown in Figure 2. The Happy and Angry classes were each increased to over 6,000 images, while the Sad class reached approximately 3,200 images. Although class imbalance remains, the augmentation increased the number of training samples for the all presenting categories. Nonetheless, the expanded dataset may help the model learn more robust decision boundaries and partially mitigate bias toward dominant classes during training.

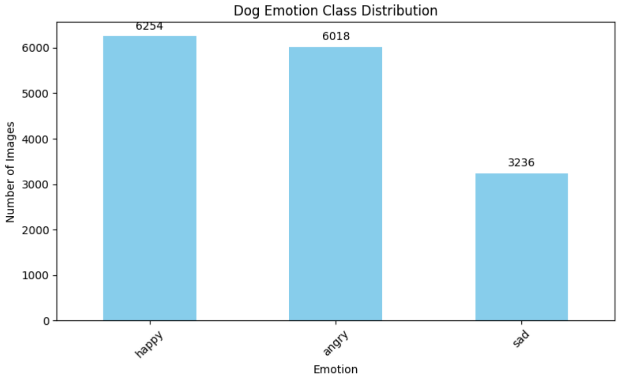


Figure 2. Post-augmentation class distribution of dog emotions

In Figure 3, an analysis of image brightness post augmentation revealed that Sad images tended to have slightly lower average brightness compared to Happy and Angry samples. This reflects inherent visual cues in the dataset or lighting characteristics preserved during augmentation. Contrast distributions were relatively consistent across all classes, suggesting that augmentation preserved tonal range without introducing class dependent artifacts. These findings suggest that while brightness may offer some weak discriminatory signal, contrast is unlikely to be a dominant feature influencing model performance.

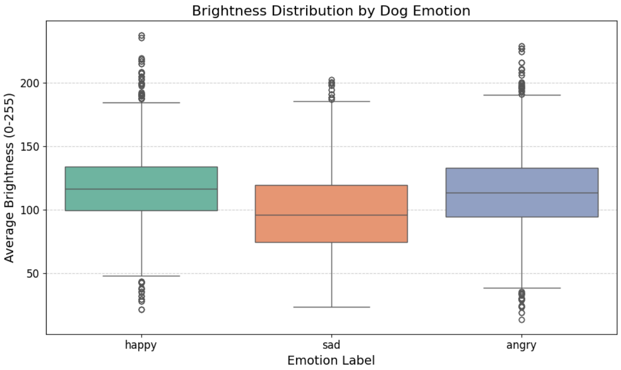


Figure 3. Brightness Distribution by Dog Emotion

As shown in Figure 4, RGB channel analysis showed that Sad images exhibited lower overall mean intensities across the all three color channels, resulting in a generally darker appearance compared to Happy and Angry. These results indicate that while color alone is not a definitive predictor of emotional label, tonal biases could influence feature learning if not carefully regularized in the model. It is also worthy to note that these patterns were observable after augmentation, suggesting that the applied transformations did not heavily disrupt the original color balance of the images.

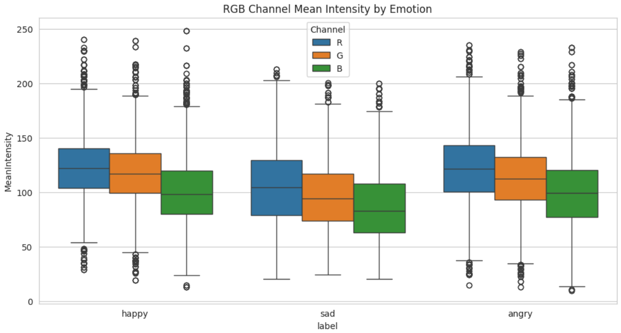


Figure 4. RGB Channel Mean Intensity by Emotion

In addition to boxplot comparisons, average RGB histograms were computed per class to capture pixel wise color distribution trends. As shown in Appendix A (Figure A1- A3), Sad Images cluster heavily in lower intensity bins across all channels, confirming their darker tonal range. Happy images display a broader RGB spread with balanced peaks, reflecting overall brightness and contrast. Angry images exhibit more mid-range red and blue intensities, indicating higher visual tension.

4. Experiments

In this section, we present a series of experiments conducted to evaluate the performance of various deep learning models in classifying dog emotions based on images. Three models were implemented and compared, a customised Convolutional Neural Network (CNN), MobileNetV2, and Resnet18. MobileNetV2 is a lightweight pre-trained architecture, while ResNet18 is a deep residual network that is known for strong feature extraction. All models were trained using a supervised learning approach, where the goal is to predict one of three discrete emotion labels (Happy, Sad, Angry) from labeled input images.

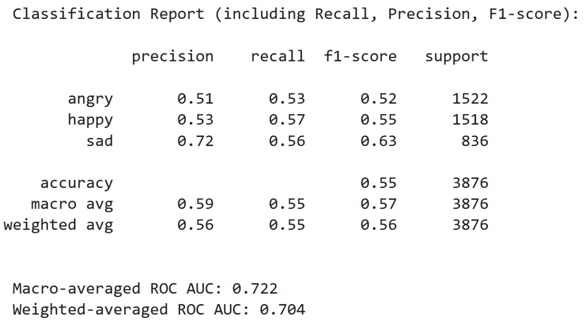
All models are trained on the same post augmentation dataset with images resized to 224 x 224 pixels. The dataset was split into an 80:20 training and validation set using stratified sampling, ensuring balanced label distribution across splits. This preprocessing pipeline was uniformly applied across all experiments to ensure comparability between model performances. Performance was assessed using standard classification metrics including accuracy, precision, recall, F-1 score and ROC-AUC.

4.1 Customized Convolutional Neural Network (CNN)

To establish a baseline, a simple Convolutional Neural Network (CNN) was constructed and trained on the augmented dataset. The architecture consists of three convolutional layers with increasing filter sizes (32, 64, 64), each followed by max-pooling to reduce spatial dimensions. The final layers include a flatten operation, a dense layer with 64 ReLU-activated units, and a softmax output layer corresponding to the number of emotion classes. The model was compiled using the Adam optimizer and categorical cross-entropy loss function, and trained for 10 epochs. Early model checkpoints were saved based on improvements in validation accuracy.

Despite its simplicity, CNN achieved a test accuracy of 55%, with macro-averaged ROC AUC of 0.722 and weighted ROC AUC of 0.704. As shown in the classification report (Figure 5), the Sad class had the highest precision (0.72) and F1-score (0.63), while the Angry and Happy classes showed more modest performance. The confusion matrix (Figure 6) further highlights the model’s key weaknesses. For instance, out of 1522 samples labeled as angry, only 808 were correctly classified, while 618 were misclassified as happy. Similarly, among happy samples, 566 were wrongly predicted as angry.

These misclassifications suggest that the CNN struggles to differentiate between visually similar emotional expressions, particularly between angry and happy, which may share overlapping facial and postural features. This limitation likely arises from the relatively shallow architecture and lack of high-level abstract representation in the CNN.



```
Classification Report (including Recall, Precision, F1-score):

              precision    recall  f1-score   support

       angry       0.51      0.53      0.52      1522
       happy       0.53      0.57      0.55      1518
         sad       0.72      0.56      0.63       836

    accuracy                           0.55      3876
   macro avg       0.59      0.55      0.57      3876
weighted avg       0.56      0.55      0.56      3876


Macro-averaged ROC AUC: 0.722
Weighted-averaged ROC AUC: 0.704
```

Figure 5. CNN Classification Report

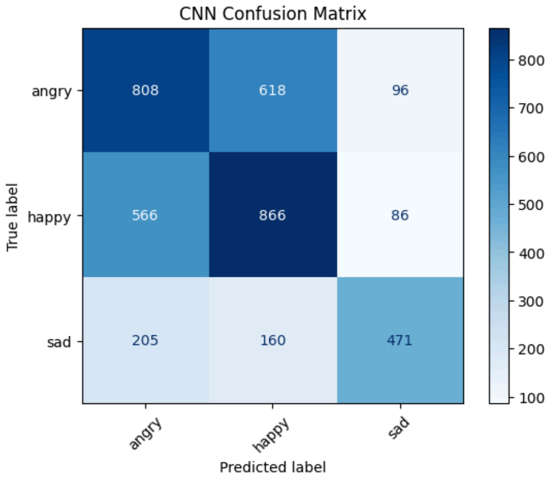


Figure 6. CNN Confusion Matrix

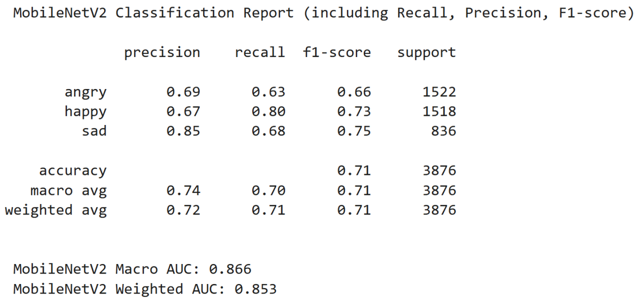
4.2 MobileNetV2

To improve baseline performance with a lightweight architecture, a MobileNetV2 network was implemented. MobileNetV2 is a compact convolutional neural network optimized for efficiency. In this setup, the pre-trained MobileNetV2 model was used as a frozen feature extractor. A global average pooling layer, dropout rate of 0.3, and a softmax classification head were added for emotion classification. The model was compiled using the Adam optimizer with categorical cross entropy loss and trained 10 epochs on the same split.

The model achieved a test accuracy of 71.0%, with macro-averaged ROC AUC of 0.866 and weighted-averaged AUC of 0.853, showing clear improvement over the baseline. Performance was strongest on the Sad class (F1 = 0.75), followed by Happy (F1 = 0.73). Although the Angry class remained the most challenging (F1 = 0.66, recall = 0.63), this marks a considerable improvement compared to the baseline CNN (F1 = 0.52, recall = 0.53).

As illustrated in the classification report (Figure 7) and confusion matrix (Figure 8), MobileNetV2 demonstrated a solid ability to differentiate between emotional categories. However, misclassifications between angry and happy remain present, suggesting some overlap in visual features. While MobileNetV2 offers strong performance with reduced computational cost, its relatively shallow architecture may limit its ability to capture fine-grained emotional nuances in complex dog expressions.

Overall, these results indicate that MobileNetV2 provides a strong trade-off between accuracy and efficiency, making it well-suited for deployment scenarios.



```
  MobileNetV2 Classification Report (including Recall, Precision, F1-score)

              precision    recall  f1-score   support

       angry       0.69      0.63      0.66      1522
       happy       0.67      0.80      0.73      1518
         sad       0.85      0.68      0.75       836

    accuracy                           0.71      3876
   macro avg       0.74      0.70      0.71      3876
weighted avg       0.72      0.71      0.71      3876


  MobileNetV2 Macro AUC: 0.866
  MobileNetV2 Weighted AUC: 0.853
```

Figure 7. MobileNetV2 Classification Report

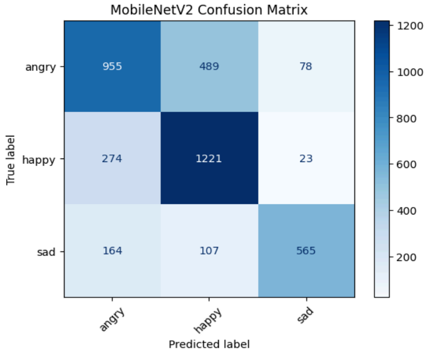


Figure 8. MobileNetV2 Confusion Matrix

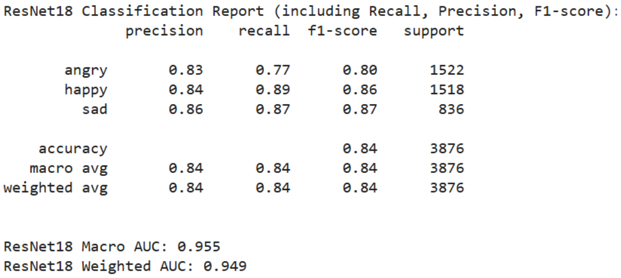
4.3 ResNet18

To further enhance classification performance, a ResNet18 architecture was employed using transfer learning. The model was initialized with pre-trained ImageNet weights, with all layers initially frozen. Fine-tuning was limited to the final convolutional block (layer4) and the fully connected layer to balance model performance and training efficiency. The dataset was stratified into training and validation sets, and early stopping was implemented to prevent overfitting. Training was performed for up to 10 epochs using the Adam optimizer and cross-entropy loss.

ResNet18 achieved the highest test accuracy among all models, reaching 84.0%, and significantly outperformed both the baseline CNN and MobileNetV2. As shown in the classification report (Figure 9) and confusion matrix (Figure 10), the model consistently delivered high precision and recall across all three emotion classes, with F1-scores exceeding 0.80. The macro and weighted ROC-AUC scores were 0.955 and 0.949 respectively, indicating strong class separability and robust generalization.

However, closer inspection of the confusion matrix reveals a notable degree of misclassification between angry and happy classes: 254 angry instances were predicted as happy, and 147 happy instances as angry. This suggests potential overlap in visual features such as open mouths, facial tension, or other subtle expressions, which the model may struggle to distinguish.

Despite these misclassifications, the model's overall performance remains strong and reliable. The observed confusion also highlights an opportunity for further refinement. Given its high performance and stable predictions, ResNet18 was selected as the backbone model for the subsequent interpretability and deployment stages.



```
ResNet18 Classification Report (including Recall, Precision, F1-score):
              precision    recall  f1-score   support

       angry       0.83      0.77      0.80      1522
       happy       0.84      0.89      0.86      1518
         sad       0.86      0.87      0.87       836

    accuracy                           0.84      3876
   macro avg       0.84      0.84      0.84      3876
weighted avg       0.84      0.84      0.84      3876


ResNet18 Macro AUC: 0.955
ResNet18 Weighted AUC: 0.949
```

Figure 9. ResNet Classification Report

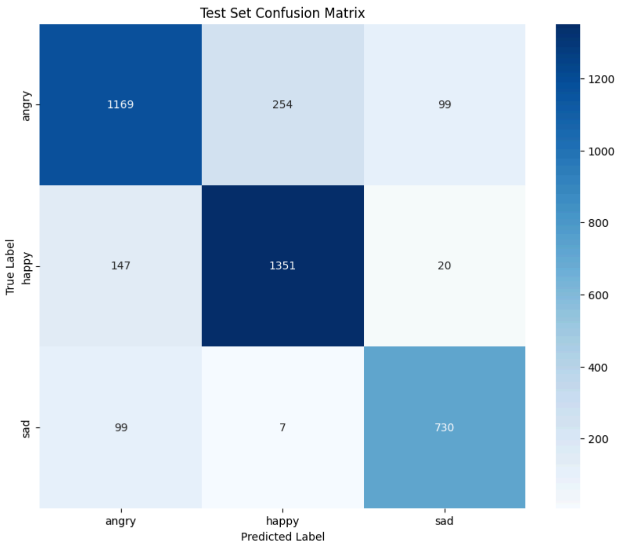


Figure 10. ResNet Confusion Matrix

5. Model Interpretability

While achieving high classification accuracy is important, it is equally critical to ensure that AI models are transparent and understandable, especially in applications involving emotion recognition. In this project, we focused on improving the interpretability of our dog emotion classification system by combining visual explanation techniques and natural language generation. Specifically, we implemented Grad-CAM to visualize the model's attention patterns and integrated an LLM to produce human-readable explanations.

5.1 Grad-CAM

Convolutional neural networks (CNNs) are powerful tools for image classification tasks, but their complex architectures often make them difficult to interpret. In our project, we aimed to address this "black box" limitation by implementing Gradient-weighted Class Activation Mapping (Grad-CAM), a technique designed to visually explain the decision-making process of CNNs.

Grad-CAM works by computing a heatmap over the input image, highlighting the regions that most strongly influence the model's classification for a given class. Specifically, it uses the gradient of the target class with respect to the final convolutional feature maps, producing a spatially resolved visualization that indicates which parts of the image were most important for the model's prediction (Melanie, 2024).

To determine the most informative layer for visualization, we compared Grad-CAM on the final convolutional blocks of two key ResNet18 layers: layer3[-1] and layer4[-1]. These are the last residual blocks within each layer sequence before the model transitions to the fully connected layers, making them well-suited for Grad-CAM, which requires access to spatially-aware activation maps. While layer3[-1] offers a higher-resolution feature map, its attention outputs were often diffuse and less focused. In contrast, layer4[-1] produced sharper, semantically meaningful heatmaps that concentrated on emotionally relevant regions of the dog's face, such as the mouth.. Based on these results, we selected layer4[-1] as the default layer for interpretability.

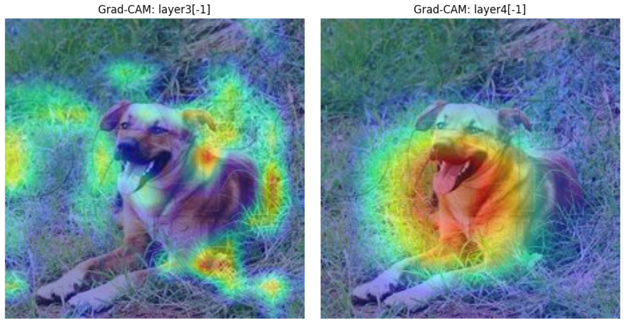


Figure 11. Comparison: Grad-CAM layer3[-1] vs. layer4[-1]

A visual example comparing both layers' outputs is provided in Figure 11. Overall, Grad-CAM enhances model transparency by helping both developers and end users understand whether the model is focusing on appropriate features, and can reveal hidden biases or failure modes, a valuable step toward responsible AI deployment (Melanie, 2024).

5.2 LLM Integration

The dog emotion detection application leverages an LLM to generate human-readable explanations for the model's predictions. After the trained ResNet model classifies the dog's emotion based on the uploaded image, the LLM is used to provide insights into the factors influencing the prediction using the Grad-CAM heatmap.

The application feeds the predicted emotion class, its probability, and information about the most influential regions of the image (derived from the heatmap) as input to the pre-trained GPT-Neo language model. The LLM then generates a natural language explanation that describes the model's reasoning in an easily understandable way.

Integrating an LLM adds value to the application by making the system's decision-making process more transparent and interpretable to users. Rather than simply outputting a predicted emotion, the LLM-generated explanations provide context about why a particular emotion was predicted. This helps build trust in the model's predictions and allows users to better understand how the AI arrives at its conclusions.

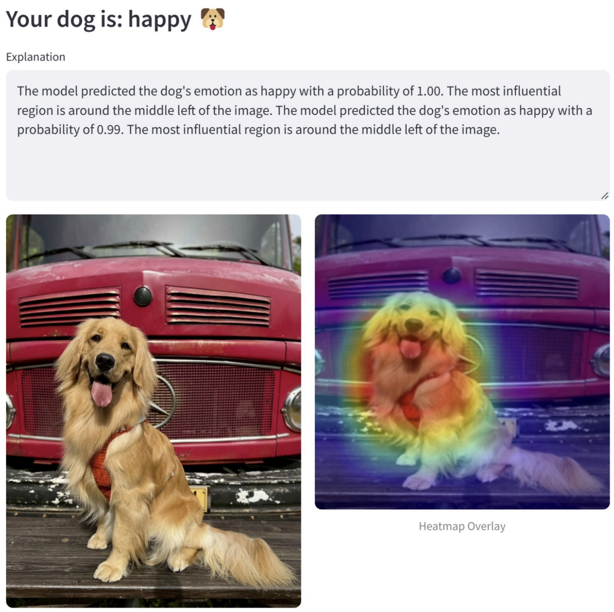
Moreover, the LLM's ability to articulate the key regions of the image that influenced the prediction (e.g. focusing on the dog's facial features) offers valuable insights into what the model is "paying attention to". This level of explainability is crucial for AI systems to be reliably used in real-world applications.

Overall, by integrating an LLM, the dog emotion detection application not only classifies emotions accurately but also generates meaningful explanations, enhancing the interpretability and trustworthiness of the AI system. This demonstrates the power of combining computer vision models with large language models to create more transparent and insightful AI applications.

5.3 Results

To demonstrate the effectiveness of our interpretability approach, we present two example outputs generated by our final model pipeline. For each case, we show the original uploaded image, the corresponding Grad-CAM heatmap highlighting the model's attention regions, and the natural language explanation generated by the LLM.

In the first example seen in Figure 12, the model correctly predicts the dog's emotion as happy with very high confidence. The Grad-CAM heatmap focuses on the central region of the image, particularly around the dog's face and open mouth, which are key visual indicators of happiness. The LLM-generated explanation accurately reflects this by emphasizing that the model attended to the middle left region, where the dog's mouth is open and tongue is visible, typical of a happy emotional state.



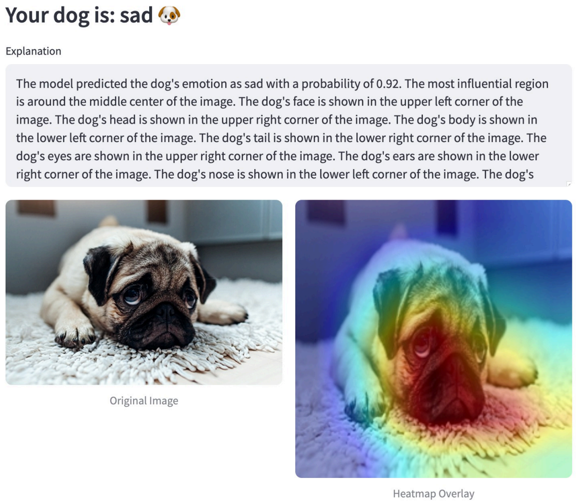
Your dog is: happy 🐶

Explanation

The model predicted the dog's emotion as happy with a probability of 1.00. The most influential region is around the middle left of the image. The model predicted the dog's emotion as happy with a probability of 0.99. The most influential region is around the middle left of the image.

Figure 12. Example 1: Happy Emotion Prediction

In the next example seen in Figure 13, the model predicts the dog's emotion as sad, with a high probability of 0.92. The Grad-CAM heatmap highlights the middle and lower parts of the image, concentrating around the dog's face and eyes. The accompanying LLM-generated explanation emphasizes that the most influential regions include the nose and eyes, which are typical indicators of sadness or low energy in canine body language.



Your dog is: sad 🐶

Explanation

The model predicted the dog's emotion as sad with a probability of 0.92. The most influential region is around the middle center of the image. The dog's face is shown in the upper left corner of the image. The dog's head is shown in the upper right corner of the image. The dog's body is shown in the lower left corner of the image. The dog's tail is shown in the lower right corner of the image. The dog's eyes are shown in the upper right corner of the image. The dog's ears are shown in the lower right corner of the image. The dog's nose is shown in the lower left corner of the image. The dog's

Figure 13. Example 1: Sad Emotion Prediction

Overall, these examples illustrate how combining Grad-CAM heatmaps and LLM-generated text explanations provides a transparent and interpretable output that supports users in understanding both the prediction and the model's reasoning process.

6. Web Interface using Streamlit

To make our model accessible, we built a user-friendly web interface using Streamlit. The interface is designed to serve users such as dog owners, veterinarians, and animal shelter staff, allowing them to upload an image and instantly receive the dog's predicted emotional state along with an explanation.

Upon image upload, the model returns the predicted emotion along with a visual heatmap using Grad-CAM from the layer4[-1] of ResNet18. This heatmap highlights the key regions that contributed to the model's prediction, such as the dog's mouth, ears, or eyes. To complement the visual explanation, we integrated an LLM to generate a short explanation describing why the model predicted that the dog is likely experiencing that emotion based on visible cues. For example, if the heatmap emphasizes the snout and the mouth is open, the LLM might explain that these features typically suggest a happy or excited emotional state.

The result is a transparent and interactive web tool that combines deep learning predictions with interpretable, user-friendly feedback that helps users understand what the model predicts and what it bases that prediction on.

7. Limitation and Future Improvement

To further improve the overall performance, interpretability, and user value of the dog emotion detection application, several key areas could be explored:

Fine-grained feature detection: A significant number of angry images were misclassified as happy (254 cases), and vice versa (147 cases), indicating overlapping visual cues such as open mouths or facial tension. Incorporating fine-grained facial feature detection (e.g., eyes, mouths, ears), localized attention modules, or even short video inputs could help the model better differentiate between these similar emotional expressions.

Expanding the range of detectable emotions: Currently, the model is limited to classifying three emotions: happy, sad, and angry. Expanding the emotion categories could make the system more informative. By collecting and labeling additional training data that captures a wider range of dog emotional states—such as fear, excitement, aggression, curiosity, or relaxation—the model could be extended to recognize more subtle emotional expressions, increasing its practical applicability in real-world pet care scenarios.

Domain-specific fine-tuning: The current LLM, GPT-Neo, is a general-purpose language model. Fine-tuning it on a dataset specifically related to dog emotions, behaviors, and visual features would likely result in more accurate and insightful explanations. By exposing the model to relevant knowledge during training, it can learn the nuances and terminology associated with interpreting canine emotions.

Incorporating additional features: Rather than relying solely on the predicted emotion class and its probability, the LLM could benefit from receiving more granular information about the dog's physical cues. Extending the computer vision model to detect specific features such as tail wagging, ear positions, facial expressions, and body posture would provide valuable context for the LLM to generate more precise explanations. These features serve as strong indicators of a dog's emotional state and can help the LLM reason about the image content more effectively.

Leveraging image captioning models: Integrating an image captioning model alongside the emotion classification model could greatly enhance the LLM's understanding of the input image. Image captioning models, such as those based on transformer architectures like VisionTransformer (ViT) or Convolutional Neural Networks (CNNs) combined with language models, can generate descriptive text summarizing the salient aspects of an image. By feeding this generated caption to the LLM, it would have access to a high-level representation of the image, enabling it to produce more coherent and contextually relevant explanations.

Human evaluation and feedback: Conducting user studies and gathering feedback from domain experts, such as veterinarians or animal behaviorists, would provide valuable insights into the quality and usefulness of the generated explanations. By incorporating human evaluation metrics and iteratively refining the LLM based on user feedback, the explanations can be made more accurate, informative, and aligned with expert knowledge.

Integrating an LLM into the dog emotion detection application has the potential to greatly enhance its interpretability and user trust. By generating human-readable explanations, the LLM can bridge the gap between the model's predictions and the end-user's understanding. However, realizing the full potential of LLMs in this context requires careful consideration of domain-specific fine-tuning, incorporation of additional features, leveraging complementary models like image captioning, employing advanced explanation generation techniques, and continuous iteration based on human evaluation and feedback. By addressing these aspects, the dog emotion detection application can provide not only accurate predictions but also meaningful and reliable explanations, ultimately improving its practical value and user experience.

8. Conclusion

In this project, we developed a deep learning-based system to classify dog emotions into three categories: Happy, Sad, and Angry, using visual data. Through the implementation of customized CNNs, MobileNetV2, and a fine-tuned ResNet18 architecture, combined with extensive data augmentation techniques, we demonstrated the feasibility of achieving robust and reliable emotion recognition in dog images despite dataset variability and emotional subtlety.

The final ResNet18-based solution delivered strong results, achieving a test accuracy of 84% and a macro-averaged AUC of 0.955. To further enhance transparency and user trust, we integrated Grad-CAM visualizations and natural language explanations generated by a Large Language Model (LLM), creating a fully interpretable and user-friendly web-based application.

Our system offers significant commercial value across several domains. In pet daycares, animal shelters, and veterinary clinics, early detection of angry emotional states enables timely intervention, reducing the risk of animal conflicts, enhancing operational safety, and improving the overall quality of care. For smart pet monitoring solutions, integrating emotion recognition allows real-time notifications, enriching the user experience and differentiating products through emotionally intelligent features.

While promising, the current system is limited to recognizing three emotional categories and relies on a general-purpose LLM, which may occasionally produce generic or less domain-specific explanations. Additionally, because the LLM bases its reasoning on indirect visual cues rather than comprehensive image understanding, there is a risk of hallucinations or inconsistencies in explanation generation.

Looking ahead, expanding the model to cover a broader range of dog emotions (e.g., fear, excitement, anxiety), integrating multimodal inputs such as video and audio signals, and fine-tuning the LLM on domain-specific datasets will be crucial to improving both classification accuracy and explanation quality. Furthermore, adapting the system to recognize emotions across other animal species presents exciting opportunities to broaden its practical applications in the fields of pet care, veterinary science, and animal welfare.

Overall, this project represents a significant step toward the development of practical, transparent, and intelligent animal emotion recognition systems that can enhance animal welfare, promote safety, and deepen human-animal connections across pet care, veterinary, smart device, and broader animal welfare industries.

References

Melanie, *What is the Grad-CAM method?* DataScientest, 2 May 2024. Available at: https://datascientest.com/en/what-is-the-grad-cam-method

Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. Journal of Big Data, 6(1), 60. https://doi.org/10.1186/s40537-019-0197-0

Mao, Y., & Liu, C. (2023). Pet dog facial expression recognition with convolutional neural network and improved whale optimization algorithm. Scientific Reports, 13, Article 30442. https://doi.org/10.1038/s41598-023-30442-0

Wu, Q., Li, J., Zhang, H., & Lin, Y. (2023). Emotion detection of dogs and cats using classification models and object detection model. International Journal of Advanced Computer Science and Applications, 14(5), 472–478. https://www.researchgate.net/publication/370729284_Emotion_Detection_of_Dogs_and_Cats_Using_Classification_Models_and_Object_Detection_Model

Lu, H., Chen, J., Liang, F., Tan, M., Zeng, R., & Hu, X. (2024). Understanding emotional body expressions via large language models. arXiv preprint arXiv:2412.12581. https://arxiv.org/abs/2412.12581

Appendices

Appendix A: RGB Channel Distribution by Emotion

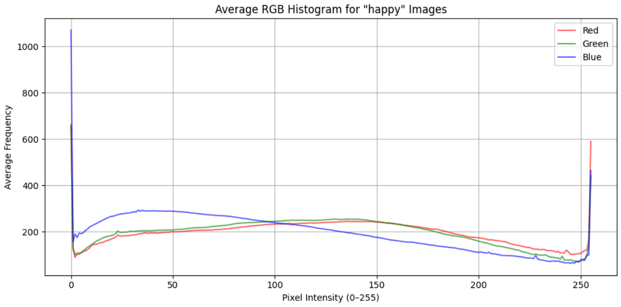


Figure A1. Average RGB Histogram for "Happy" images.

This figure shows the mean pixel intensity distribution across Red, Green, and Blue channels for all images labeled as *Happy* after augmentation.

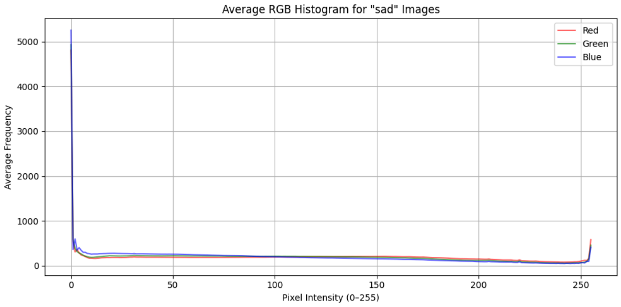


Figure A2. Average RGB Histogram for "Sad" images.

This plot illustrates the color distribution for *Sad* class images, where lower intensity values dominate, particularly in the blue channel.

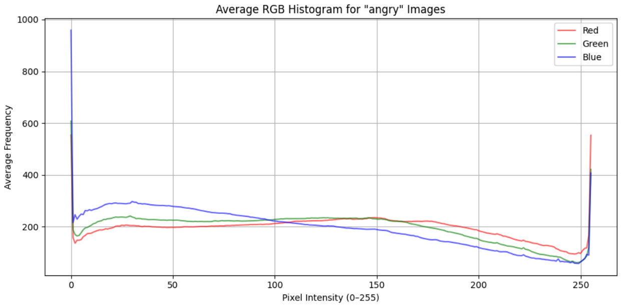


Figure A3. Average RGB Histogram for "Angry" images.

Displays the RGB intensity trends for *Angry* class, showing relatively warmer tones with elevated red and green values compared to *Sad*.