Comparison of a CNN and a Vision Transformer for Classifying Basic Cat Facial Expressions

Izhar Octrafilian Susilo
Computer Science Department
School of Computer Science
Bina Nusantara University
Jakarta, Indonesia
izhar.susilo@binus.ac.id

Nicholas Andrew Marvell
Computer Science Department
School of Computer Science
Bina Nusantara University
Jakarta, Indonesia
nicholas.marvell@binus.ac.id

Henry Lucky
Computer Science Department
School of Computer Science
Bina Nusantara University
Jakarta, Indonesia
henry.lucky@binus.ac.id

Derwin Suhartono
Computer Science Department
School of Computer Science
Bina Nusantara University
Jakarta, Indonesia
derwin.suhartono@binus.edu

***Abstract*—This electronic document is a "live" template and already defines the components of your paper [title, text, heads, etc.] in its style sheet. **CRITICAL: Do Not Use Symbols, Special Characters, Footnotes, or Math in Paper Title or Abstract.* (*Abstract*)**

Keywords—component, formatting, style, styling, insert (*key words****)***

I. INTRODUCTION (*HEADING 1*)

Recognizing facial expressions in pets, like cats, is one of the interesting and a growing area in Artificial Intelligence (AI). Understanding a cat's emotion allows us to make improvements and knowing the accuracy of a cat's expressions can be beneficial in many applications, such as for the pet owners, the veterinarians, and for animal welfare groups. However, knowing and identifying cats' facial expressions is really difficult because their faces have different patterns in the fur and small changes that are also hard to notice.

To identify their expressions, we use image recognition, the most common method used are Convolutional Neural Networks (CNNs) being the most widely implemented. Previous studies **[1] [3]** have shown that CNNs can successfully classify a cat's facial expressions, and their accuracy can be improved with more training. The studies utilized CNN models trained on ImageNet and Stanford Dogs dataset, achieving high accuracy in pet emotion recognition, but performance was inconsistent for different breeds. Another study **[4]** examined 1000 cat facial images and found that CNN models become more accurate with repeated training sessions. Additionally ,studies **[9]** showed that CNN-based landmark models analysing cat facial landmarks achieved higher accuracy compared to texture-based methods. However, CNNs are mainly focussing on small details like edges and textures, which might not be enough to fully understand the cat's emotions.

But recently, there has been a new method called Vision Transformers (VITs) that has become more popular. VITs look at an image in a different way, they do it by breaking the images into smaller parts and understanding how they relate to each other. Research studies in **[6][7]** showed that ViTs are good at recognizing emotions, especially in complex images. The studies compared the results of CNN and ViT and found that comparing CNN and ViT-based models on facial expression datasets (JAFFE, CK+) reported that ViTs achieved higher accuracy in recognizing nuanced facial expressions. Moreover, studies **[8**], have shown that CNNs tend to perform better on small datasets, whereas ViTs excel with larger datasets. Despite this, limited research has been conducted on comparing CNN and ViT specifically for recognizing animal expressions, particularly in cats. Additionally, the impact of dataset size on performance differences between CNN and ViT remains unclear.

Our research is aiming to find the comparison between CNN models (ResNet18, VGG16) and ViT models (DINOv2, BeiT) for classifying basic cat facial expressions. Between those two methods, we want to find which performances using two datasets, one small-scale and one large-scale datasets. By doing this comparison, we want to find which approach is more effective in recognizing cat facial expressions and how datasets size influence their accuracy. These results of this research can make a contribution for AI, for improving understanding cat emotions that can benefit for pet owners and for those veterinarians.

II. RELATED WORKS

Emotion recognition using artificial intelligence has been studied widely including on human facial expression and animal expressions. One of the methods often used is Convolutional Neural Networks (CNN). CNN is considered the traditional method, however there is a new and upcoming method called Vision Transformers (ViTs). CNN works by processing an image through several layers, while ViTs process them by splitting them into patches. Researchers are shifting to ViTs as Vision Transformers

offer greater ability to capture global relationships within an image.

When using CNN, a study **[3]**, utilizes CNN features to unravel the intricate language embedded in the facial expressions of cats. The study used a dataset with a total of 1000 photos for training and 300 for validation. The study shows that the model's accuracy improves every iteration (Epoch) done. Epoch represents one complete pass of the training dataset through the model which allows it to refine its learning over time. It shows that for the first iteration, the accuracy is only at 55.56%, but over time, it improved, and eventually reached 95.71% by the 10th iteration.

Another study using CNN **[1]** utilizes Faster-RCNN. Faster-RCNN differs from CNN as it detects regions of interest before classifying emotions which optimizes its ability to analyze both facial expression and body language in pets. The study trained its model on various datasets which enabled the model to differentiate happiness, fear, and neutrality. The study specifically underlined the difficulty in emotion classification as there were a lot of differences in facial anatomy and non-uniform emotional cues across species. The study showed that Faster-RCNN successfully classified both pet species and emotion in a high accuracy.

An experiment **[5]** used CNN and Improved Whale Optimization Algorithm to recognize the facial expression of dogs. Using a dataset that includes 315 pet dog images labelled with 5 emotions(normal, happy, sad, angry, fear). Results show that the integrated CNN-IOWA model demonstrated a significant improvement compared to standard CNN models. In a study **[9]** also uses CNN in recognizing domestic cats facial expression, with a result of 98% overall.

For ViT (Vision Transformers) that has been gaining attention to handle complex datasets, theres has been a study **[7]**, that explains how ViT can recognize facial expressions, unlike the CNN which they process an image by scanning the local patterns, but with ViT they break images into patches and use self attention to figure out relationships between different parts of the image.This allows them to pick up on subtle emotional details, even in noisy or complicated pictures. The study found that ViTs did particularly well at recognizing emotions that aren't super obvious or require understanding the full context of the image.

The study **[6]**, evaluated thirteen Vision Transformer (ViT) models on facial emotion recognition tasks using the RAF-DB, FER2013, and a newly balanced dataset. findings indicate that Mobile ViT and Tokens-to-Token ViT models were the most effective, followed by PiT and Cross Former models. The paper shows how ViT performs exceptionally well in recognizing the emotions in noisy and complex images.

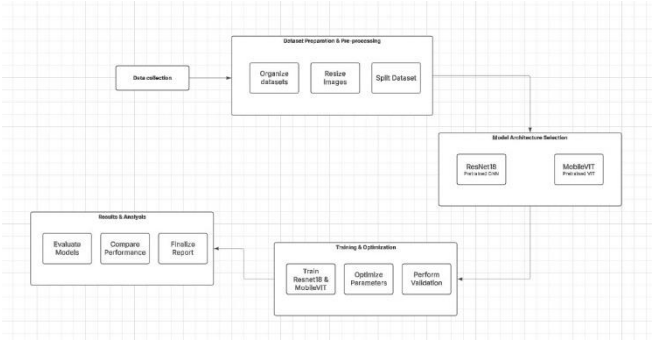
But how do we compare CNN and ViT when it comes to human expression or animal expression? There's this study **[8]**, where they look at the strengths and weaknesses between those 2 methods, now CNN is great at picking up local features like the edges, textures, or small details which makes them more effective for smaller datasets. ViT is better at capturing the bigger picture and handling large datasets which makes them a strong option for something like our research paper which is about Classifying Basic Cat Facial Expressions. But still CNN works better when there is a limited dataset and easier to train.

III. METHODOLOGY

METHODOLOGY

A. Research Design

This research employs a quantitative, comparative design to evaluate the performance of two deep learning models, ResNet18 (CNN) and MobileViT (Vision Transformer), in classifying cat facial expressions. The primary goal is to determine which model achieves higher accuracy in categorizing cat images into one of four emotional states: "Pleased," "Angry," "Alarmed," and "Calm."



B. Dataset

1. Source: The dataset will be sourced from the "Domestic Cats facial expressions Computer Vision Project" on Roboflow Universe, containing approximately 500 pet images

2. Image Selection and Emotion Labeling: A subset of the images (approximately 250 images) will be selected based on image quality and diversity. Each selected image will be manually labeled with one of the four target emotions ("Pleased," "Angry," "Alarmed," "Calm").

3. Dataset Partitioning: The labeled dataset will be partitioned into three subsets:

   - Training set (70%): Used for training the models.
   - Validation set (15%): Used for hyperparameter tuning and monitoring model performance during training to prevent overfitting
   - Test set (15%): Used for final evaluation of the trained models.

C. Model Architecture

1. CNN: ResNet18

   - A pre-trained ResNet18 model will be used

- The final fully connected layer will be modified to output four classes.

2. Vision Transformer: MobileViT
    - The MobileViT architecture will be employed.
    - Pre-trained weights will be used if available, with the classification head adjusted for four classes.

D. Training Procedure

1. Implementation:
    - Models will be implemented using PyTorch.
    - Pre-trained weights will be loaded to leverage transfer learning.
2. Optimization:
    - Use the Adam optimizer for both models.
    - Apply cross-entropy loss as the loss function.
3. Training Process:
    - Train both models on the training set for a fixed number of epochs (e.g., 10-20).
    - Monitor validation loss to prevent overfitting using early stopping.
4. Evaluation Metrics:
    - Accuracy, Precision, Recall, F1-Score, and Confusion Matrix will be used to evaluate model performance on the test set.

E. Analysis

1. Comparative Performance: Compare evaluation metrics for ResNet18 and MobileViT.
2. Statistical Analysis: Conduct tests (e.g., t-tests) to determine if performance differences are statistically significant.
3. Qualitative Analysis: Analyze confusion matrices to identify which emotion categories are most challenging for each model.

REFERENCES

[1] Sinnott, R. O., Aickelin, U., Jia, Y., Sinnott, E. R. J., Sun, P. Y., & Susanto, R. (2021). Run or Pat: Using Deep Learning to Classify the Species Type and Emotion of Pets. *2021 IEEE Asia-Pacific Conference on Computer Science and Data Engineering, CSDE 2021, December 2021*. https://doi.org/10.1109/CSDE53843.2021.9718465

[2] Feighelstein, M., Henze, L., Meller, S., Shimshoni, I., Hermoni, B., Berko, M., Twele, F., Schütter, A., Dorn, N., Kästner, S., Finka, L., Luna, S. P. L., Mills, D. S., Volk, H. A., & Zamansky, A. (2023). Explainable automated pain recognition in cats. *Scientific Reports*, *13*(1), 1–16. https://doi.org/10.1038/s41598-023-35846-6

[3] Jain, P., Pandey, A. K., Manoj, V., Pandey, K., & Yadav, B. (2023). *Facial Emotion Recognition of Cat Breeds by Using Convolution Neural Network*. *7*(6), 58–63.

[4] Bhagat, D., Vakil, A., Gupta, R. K., & Kumar, A. (2024). Facial Emotion Recognition (FER) using Convolutional Neural Network (CNN). *Procedia Computer Science*, *235*(2023), 2079–2089. https://doi.org/10.1016/j.procs.2024.04.197

[5] Mao, Y., & Liu, Y. (2023). Pet dog facial expression recognition based on convolutional neural network and improved whale optimization algorithm. *Scientific Reports*, *13*(1), 1–20. https://doi.org/10.1038/s41598-023-30442-0

[6] Bobojanov, S., Kim, B. M., Arabboev, M., & Begmatov, S. (2023). Comparative Analysis of Vision Transformer Models for Facial Emotion Recognition Using Augmented Balanced Datasets. *Applied Sciences (Switzerland)*, *13*(22). https://doi.org/10.3390/app132212271

[7] Chaudhari, A., Bhatt, C., Krishna, A., & Mazzeo, P. L. (2022). ViTFER: Facial Emotion Recognition with Vision Transformers. *Applied System Innovation*, *5*(4). https://doi.org/10.3390/asi5040080

[8] Maurício, J., Domingues, I., & Bernardino, J. (2023). Comparing Vision Transformers and Convolutional Neural Networks for Image Classification: A Literature Review. *Applied Sciences (Switzerland)*, *13*(9). https://doi.org/10.3390/app13095521

[9] Abubakar Ali, Onana Oyana, C. L. N., & Salum, O. S. (2024). Domestic Cats Facial Expression Recognition Based on Convolutional Neural Networks. *International Journal of Engineering and Advanced Technology*, *13*(5), 45–52. https://doi.org/10.35940/ijeat.e4484.13050624

[10] Wu, Y. (2023). Emotion Detection of Dogs and Cats Using Classification Models and Object Detection Model Emotion Detection of Dogs and Cats Using Classification Models and Object Detection Model. May.