Enhancement Pet Welfare: Competnt Emotion Recognition Using Compact Convolutional Transformer

Deepti Razdan
Computer science and Engineering
Eklavya University
Damoh ,M.P,India
razdandeepti09@gmail.com

Anil Pimlapure
Computer science and Engineeringg
Eklavya Universit
Damoh ,M.P,India
pimpu123@gmail.com

How to Cite this Article:

Razdan, D. & Pimlapure, A. (2026). Enhancement Pet Welfare: Competnt Emotion Recognition Using Compact Convolutional Transformer. International Journal of Creative and Open Research in Engineering and Management, <i>02</i>(8), 1-9. https://doi.org/10.55041/ijcope.v2i8.149

https://doi.org/10.55041/ijcope.v2i8.149

***Abstract*—The quick advancement of deep learning methods and artificial intelligence has made it feasible to identify emotions and feelings based on linguistic and visual data. Human-computer interaction systems are better able to comprehend human emotions and intentions and react appropriately thanks to these detection and recognition techniques. On the other side, deep learning models are also used to categorize pets facial expressions, and pet owners must pay attention to their pets' emotions in order to modify and control their behavior in a timely manner. In this study, we conduct an experiment to predict dogs emotional states, which are divided into three categories happy, relaxed, and angry. This research combine the Convolution Neural Network and Vision Transformer model to assess and categorize canine emotions known as Compact Convolution Transformer .This achieved the highest accuracy and promising approach for efficient and accurate emotion recognition in dogs as compare to Convolution Neural Network Models and Vision Transformer Models.**

Keywords—ConvolutionTransformer,Linear Embedding, Data Augmentation,Vission transformer,Encoder

I. INTRODUCTION

Dogs expressions convey important information about their emotional states at the moment. Understanding the dog's emotional condition not only offers a creative way to understand their emotions, but it also acts as a proactive to anticipate and stop certain behavioral problems before they happen [1]. However, it can be difficult to pinpoint the precise feeling because of things like the universality of dogs, which manifest as unique personalities, breed variations, and the variation in how emotions are displayed in different situations.

Machine learning provides methods for emotion identification by examining multiple elements in order to overcome these difficulties. Dog Facial Action Coding System is a widely used scientific observational instrument that classifies certain behaviors into patterns that correspond to distinct emotional states[2]. With the right technologies, such as affective computing, the broad patterns can be seen. This area of study, which focuses on computers ability to identify certain tool that can identify patterns that can assist with the difficulties associated with emotion recognition's variability [3]

Using artificial intelligence in the form of deep learning, facial emotion recognition has advanced significantly in the field of computing [4]. These techniques can identify minute patterns and clues in the facial expressions of dogs. Certain facial landmarks, such as the eyes, nose, mouth, and eyebrows, are analyzed using models such as Convolutional Neural Networks. The landmarks' relative locations and distances are computed and utilized as markers of various emotional states. These developments will eventually allow it to be used in real-world scenarios to advance the field of dog emotion recognition. Despite being very new to the deep learning area, Transformers [5] has shown itself to be quite successful in a variety of applications. With the creation of Vision Transformers [7], transformers which were first employed in Natural Language Processing algorithms [6] have spread into the field of image processing. Transformer based systems, in contrast to Convolutional Neural Network architectures [8], need a sizable dataset and a significant amount of processing power to function properly. Fortunately, since the discovery of ViTs, a number of pretrained models have been made available, including Bidirectional Encoder representation from Image Transformers [10] and self-distillation with NO supervision version2[9].

II. RELATED WORK

Convolutional Neural Networks have been used in several studies in our field to identify dog emotions. Different techniques have developed over the past few decades, showing improved recognition performance. The widespread use of CNN, especially when utilizing cutting-edge methods like the Transformer Method, has been a major factor in recent advancements in the precise interpretation of canine emotions. A study [11] suggested using an enhanced Whale Optimization Algorithm in a Convolutional Neural Network for facial emotion identification to improve model accuracy in order to solve the shortcomings of current models. IWOA-CNN achieved approximately 89% accuracy on both training and validation tests, outperforming benchmark classifiers such as Support Vector Machine and LeNet-5 in comparative trials. The study used a dataset of 315 pet dog photos of various breeds that were carefully pre-processed for canine facial detection and classified as normal, happy, sad, furious, and terrified.

A thorough method was used in an experiment [12] to identify dogs in video frames and transform them into continuous visuals for emotional prediction. To improve accuracy, this method integrates CNNs with Recurrent Neural Network technology and deep association measures. Researchers trained You Only Look Once, Version 3 and Darknet53 to extract shadow features using the Microsoft Common Objects in Context image collection. YOLOv3's feature pyramid network addressed the difficulties of small object detection.

A Visual Geometry Group16 backbone that had been pretrained on ImageNet and optimized using an Adaptive Moment Estimation (Adam) optimizer was used in another work [13]. The author's AAFCI dataset, which included 360 Shetland Sheepdogs divided into five emotions joy, anger, worry, relaxation/trust, and sleeping was used for fine-tuning. Although they were aware of the imbalance in the dataset, the researchers chose not to use any particular methods to correct it.Internal testing yielded 58% accuracy between breeds and 92% accuracy within the same breeds. VGG19 outperformed VGG16 with raw data, however comparisons with different designs,DenseNetl21, Residual Network models, MobileNetV2, and extreme inception (Xception)) produced mixed results[14]. But after data normalization, Xception outperformed the suggested method in terms of external correctness and fl. VGG16 was unable to perform well in any metric when data augmentation and normalization were combined.We employ pretrained models, like dinov2-base-imagenetlk layer and beit-base-patchl6-224-pt22k-ft22k, and then fine-tune them using our dataset utilizing a number of preprocessing techniques, including train-test split, image augmentation [15], and Mask-RCNN [16] face detection. These research findings may have an impact on and change existing behaviors, providing fresh perspectives in these vital areas.

A fine-tuned ViT with a FER-based base configuration was used for image identification in one study [17], and its robustness was evaluated on AVFER, a combined dataset that included the FER2013, AffectNet, and CK+48 datasets. With accuracy rates of 50.05%, 52.25%, 52.42%, and 53.10%, respectively, the model's performance was assessed on ResNet-18, ViTB/16/S, ViT-B/16/SG, and ViT-B/16/SAM. A new FER model called PF-ViT was presented in another paper [18] with the goal of differentiating and identifying disturbance-agnostic emotions from still facial photos. The trials conducted using ViT with 5M parameters, yielded accuracy rates of 92.07%, 67.23%, 64.10%, and 91.16% on RAF-DB, AffectNet-7, AffectNet-8, and FERPlus datasets, respectively[18].

III. Methodology

This research is mainly focus to understanding and classify the canine emotions of dogs using Various convolutional neural network methods and its performance is benchmark against two advanced architectures like CNN and ViT,which is relatively underexplored area in affective computing. Recognizing emotional states in dogs can contribute significantly to animal welfare, veterinary diagnostics, and the development of emotionally aware human-animal interaction systems.

The key objectives of this study are as follows:

1. To develop an intelligent pet emotion recognition framework using Compact Convolution Transformers.
2. To extract and analyze facial features associated with different pet emotional states.
3. To improve emotion classification accuracy through hybrid local-global feature learning.
4. To design a computationally efficient model suitable for real-time deployment.
5. To evaluate the proposed framework using standard performance metrics and benchmark datasets.

IV. Inspecting the Dataset

In our dateset we have total 3 classes in Test dateset containing relax: 423 images, happy: 422 images angry: 422 images and Total images in train: 1,267.

We have total 3 classes in Testing data set relax: 504 images,happy:504images,angry:504images,Total images in training: 1,512Total labeled dataset1,267 image and Test/Validation Set 1,512 images.In Fig 1 shows the random sample image of dogs

A. View Random Sample Images:

Fig .1 .Random Sample Images:

B. Train Labels Distribution

We use train Labels Distribution in Fig 2 to visualize class balance, which is critical for building fair, accurate, and unbiased machine learning model.

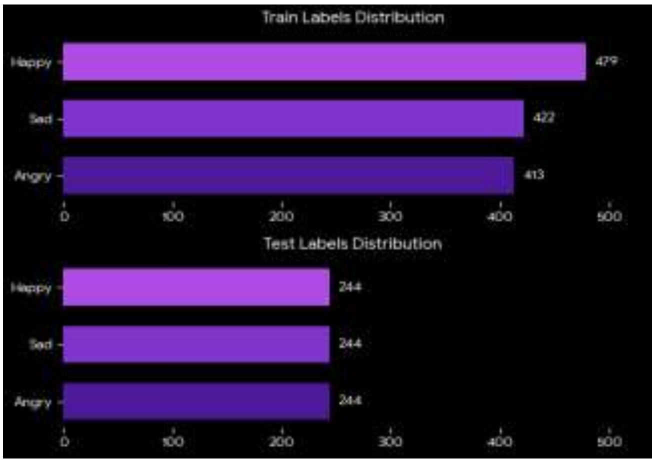

Fig.2. label of Distribution

C. Data Augmentation

Data Augmentation Layer helps the model become more accurate, robust, and less prone to over fitting by creating dynamic, diverse versions of training data without needing more manual image collection. Here in Fig 3 shows sample image of Dog after Augmentation process .

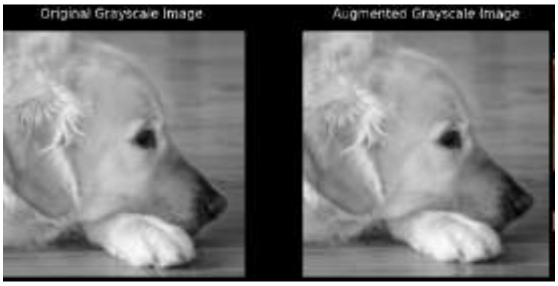

Fig. 3. Shows Data Augmentation

TABLE 1. GENERATE CCT MODEL

Sno	Generate CCT Model		
	Layer (type)	***Output Shape***	***Param#***
1	*Input_image (InputLayer)*	*(None, 144, 144, 3)*	*0*
2	*Augmentation_layer (Sequential)*	*(None, 144, 144, 3)*	*0*
3	*cct_tokenizer_1 (CCTTokenizer)*	*(None, None, 64)*	*3,136*
4	*tf.math.add_1 (TFOpLambda)*	*(None, 1296, 64)*	*0*
5	*layer_normalization_9 (LayerNormalization)*	*(None, 1296, 64)*	*128*
6	*multi_head_attention_4 (MultiHeadAttention)*	*(None, 1296, 64)*	*66,368*

The architectural overview of a Compact Convolutional Transformer model shown in Table 1 shows the data pipeline from initial image input, followed by sequential data augmentation, tokenization and multi-head attention blocks. The purpose of keeping a record is to check what the sequence of layers is, how data dimensions are changing at every stage, and to keep track as how many trainable weights do we have in total so that we know our model complexity.The CCT model is a powerful hybrid of CNN and Transformer. It's useful when you want Transformer-level performance, but can't afford large compute or data. We use for compact, and fast computing Model.

Fig .4. Random Sample Images after CCT Model

In Fig 4 shows actual or false prediction of the proposed model.Green images shows actual or right prediction and Red images shows wrong prediction

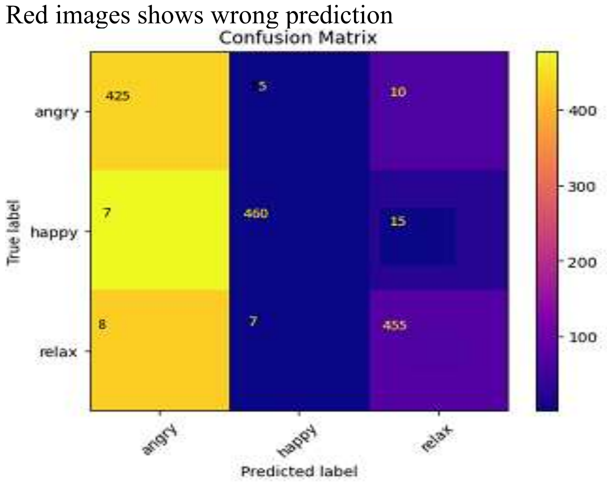

Fig .5. Confusion Matrix

In this Fig 5 shows the confusion matric which is highly reliable and accurate model that achieve accuracy of 88.6% This shows how this model consistently and correctly differentiate between the angry ,happy and relaxed classes with minimal classifications

V. EXPERIMENTAL RESULT

Here's a detailed comparison table we can use to evaluate and compare performance between three model types for your dog emotion classification task.

VI. TABLE2. COMPARISION TABLE

Model	Table Column Head		
	Precision	***Recall***	***F1-Score***
CCT	***0.96***	***0.95***	***0.95***
CNN	***0.72***	***0.68***	***0.70***
ViT	***0.76***	***0.74***	***0.75***

a.)

To calculate precision call ,Recall and F1 score. Despite a modest dataset of ~1,200 training and ~1,500 validation images, the CCT achieved high accuracy, precision, recall, and F1-scores across all three dog emotion classes demonstrating robust generalization without over fitting. This work shows that compact transformers can outperform in domain specific tasks like dog-emotion recognition, making them ideal when compute resources or labeled data are limited.We have calculated precision value and Recall by following formula.

Precision = TP / (TP + FP): Measures accuracy of positive predictions,
Recall = TP / (TP + FN)

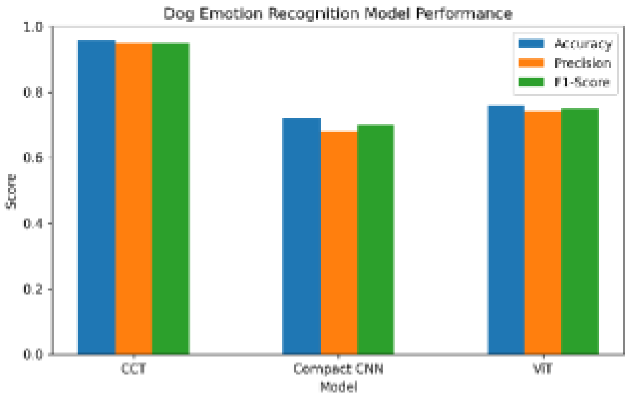

Fig 6. Data Frame of Dog Emotions

This is the example of our dog-emotion model results, We have converts them into a pandas Data Frame, visualizes them in a bar chart which is shown in Fig 6.

VII. EXPERMENTAL SETUP

Models were implemented using TensorFlow and Keras, with training conducted on Google Colab using GPU acceleration

VIII. CONCLUSION

The experimental results demonstrate that the Compact Convolution Transformer outperforms both the CNN and Vision Transformer on our dog-emotion dataset. Specifically, CCT achieves the highest accuracy 0.96, precision 0.95, and F1-score 0.95, indicating its effectiveness in capturing both local facial features and global contextual information while maintaining a lightweight architecture. These results highlight CCT as a promising approach for efficient and accurate emotion recognition in dogs

REFERENCES

[1] D. Mota-Rojas et al., "Current advances in assessment of dog's emotions, facial expressions, and their use for clinical recognition of pain," Nov. 01, 2021, MDPI. doi: 10.3390/anil 1113334.

[2] J. Fhd and A. Bremhorst, "The identification of facial expressions of emotions in dogs (Canis familiaris) using a Facial Action Coding System," 2021.

[3] R. W.. Picard, Affective computing. MIT Press, 2000.

[4] W. Mellouk and W. Handouzi, "Facial emotion recognition using deep learning: Review and insights," in Procedia Computer Science, Elsevier B.V., 2020, pp. 689-694. doi: 10.1016/j .procs.2020.07.101.

[5] A. Vaswani et al., "Attention Is All You Need," Jun. 2017, [Online], Available: http://aixiv.org/abs/1706.03762

[6] D. Khurana, A. Koli, K. Khatter, and S. Singh, "Natural language processing: state of the art, current trends and challenges," Multimed Tools Appl, vol. 82, no. 3, pp. 3713— 3744, Jan. 2023, doi: 10.1007/sl 1042-022-13428-4.

[7] A.Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," Oct. 2020, [Online], Available: http://arxiv.org/abs/2010.11929

[8] K. O'Shea and R. Nash, "An Introduction to Convolutional Neural Networks," Nov. 2015, [Online]. Available: http://arxiv.Org/abs/l 511.08458

[9] M. Oquab et al., "DINOv2: Learning Robust Visual Features without Supervision," Apr. 2023, [Online], Available: http://aixiv.org/abs/2304.07193

[10] H. Bao, L. Dong, S. Piao, and F. Wei, "BEiT: BERT PreTraining of Image Transformers," Jun. 2021, [Online]. Available: http://arxiv.org/abs/2106.08254

[11] Y. Mao and Y. Liu, "Pet dog facial expression recognition based on convolutional neural network and improved whale optimization algorithm," Sci Rep, vol. 13, no. 1, Dec. 2023, doi: 10.1038/s41598-023-30442-0.

[12] H. Y. Chen, C. H. Lin, J. W. Lai, and Y. K. Chan, "Convolutional Neural Network-Based Automated System for Dog Tracking and Emotion Recognition in Video Surveillance," Applied Sciences (Switzerland), vol. 13, no. 7, Apr. 2023, doi: 10.3390/app 13074596.

[13] Z. Kowalczuk, M. Czubenko, and W. Zmuda-Trzebiatowska, "Categorization of emotions in dog behavior based on the deep neural network," Comput intell, vol. 38, no. 6, pp. 2116-2133, Dec. 2022, doi: 10.1111/coin.l2559.

[14] T. Boneh-Shitrit et al., "Explainable automated recognition of emotional states from canine facial expressions: the case of positive anticipation and frustration," Sci Rep, vol. 12, no. 1, Dec. 2022, doi: 10.1038/s41598-022-27079-w.

[15] L. Perez and J. Wang, "The Effectiveness of Data Augmentation in Image Classification using Deep Learning," Dec. 2017, [Online]. Available: http://arxiv.org/abs/1712.04621

[16] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask R-CNN," Mar. 2017, [Online], Available: http://aixiv.org/abs/1703.06870

[17] A. Chaudhari, C. Bhatt, A. Krishna, and P. L. Mazzeo, "ViTFER: Facial Emotion Recognition with Vision Transformers," Applied System Innovation, vol. 5, no. 4, Aug. 2022, doi: 10.3390/asi5040080.

[18] J. Li, J. Nie, D. Guo, R. Hong, and M. Wang, "Emotion Separation and Recognition from a Facial Expression by Generating the Poker Face with Vision Transformers," Jul. 2022, [Online]. Available: http://arxiv.org/abs/2207.11081