Pre-Trained Multi-Modal Transformer for Pet Emotion Detection

Run Guo
Faculty of Computer Science, Dalhousie University, Halifax, Canada

Keywords: Emotion Detection, Pet's Facial Expression, Multi-Modal Transformer, Feature Representations, Deep Learning.

Abstract: With the rapid development of artificial intelligence and deep learning technology, emotion recognition and emotion detection based on visual information and language information become possible. These recognition and detection methods help to better understand human emotions and intentions in human-computer interaction systems and respond accordingly. On the other hand, more families with pets need to pay attention to the pet's emotions, so as to adjust and manage the pet's behaviour in time, and the deep learning model is also used to classify the pet's facial expressions. This paper proposes a pre-trained multi-modal transformer emotion detection system, which is first pre-trained on a human emotion detection dataset including speech and facial expression data, and then takes the labelled animal voice and expression data as small-sample task data, This approach utilizes an unlabelled corpus for pre-training, which meets the requirements of adequately training model parameters and preventing model overfitting, and finally uses representations of these models for few-shot tasks. Experimental results on video datasets show that the proposed multimodal transformer emotion detection system has good classification results on video datasets containing both sound and visual information.

1 INTRODUCTION

Humans have kept pets for thousands of years, and many pets have become members of their families. The pet economy around pets is becoming more and more prosperous. The growing number of young singles and seniors living alone in cities has also contributed to the rapid development of the pet economy. During Covid19, such living alone and being quarantined increased compared to before, and in line with this, global pet, dog and cat adoptions peaked in the early epidemic stages of the pandemic (Ho, Hussian, & Sparagano, 2021). Since felines are more adapted to family life, human families have relatively more cats.

Some studies have shown that in animals, especially mammals, there are also human-like emotional responses such as fear, and animals will show some behavioural and communication changes when these emotions appear (Bennett, Gourkow, & Mills, 2017). In addition to verbal expression, the way of judging human emotions can also rely on facial expressions. Studies on animals such as mice and cats have shown that the relationship between facial expressions and emotions not only exists in humans, but also widely exists in the faces of various animals. Since animals cannot express in rich language, their keepers can try to infer the animal's current mood by judging the animal's facial expressions.

With the rapid development of computer vision-related fields, researchers believe that emotion detection also has good application prospects in these fields. Human-related emotion detection is widespread, such as emotion detection in fields such as education and medical care. In the field of games and websites, animal-based emotion detection, such as the pain level of animals, also has good application prospects. For example, using methods such as SVM to classify the facial expressions of sheep can complete the automatic emotional evaluation of sheep (Singh et al, 2020).

In addition, researches have carried out emotion detection and analysis for dog barking. By extracting the characteristics of dog barking to determine the emotional state of the dog, it helps breeders to better identify the dog's emotion, and through interaction with the dog to cultivate feelings, help Humans build good pet relationships. Generally speaking, pet emotions can be divided into six types, or can be distinguished by two-dimensional models. Although

Guo, R.
Pre-Trained Multi-Modal Transformer for Pet Emotion Detection.
DOI: 10.5220/0011961500003612
In *Proceedings of the 3rd International Symposium on Automation, Information and Computing (ISAIC 2022)*, pages 574-579
ISBN: 978-989-758-622-4; ISSN: 2975-9463

there have been some emotional speech datasets based on human speech data, but few datasets for animal emotions, some researchers have developed a way to define cat emotions in three emotional states (Shou, Xu, Jiang, Huang, & Xiao, 2021).

Animal sound datasets generally come from online videos, with many breeders sharing their interactions with animals. In the construction of this type of data set, it is found that it is impossible to fully judge emotions through a small number of animal calls. Although video explanations and labels are often included in online videos, there are often background noises such as music in these sound data. Music is removed; in addition to using sounds to express emotions, animals also use body movements and facial expressions to express their emotions.

The rapid development of artificial intelligence and its machine learning technology has provided many new methods and means for solving the problem of emotion recognition, especially in the field of computer vision (Sinnott et al, 2021). Machine learning makes the traditional manual selection of suitable features develop into automatic learning of suitable features. The main basis is that the multi-layer neural network in deep learning can learn abstract features from low-level to high-level level by level to represent the information in the computer vision carrier. For example, an algorithmic competition for cat and dog classification emerged in Kaggle competitions.

To sum up, there are few studies in the field of animal emotions, and most of the studies have to be based on the data and recognition methods of human emotion recognition. The recognition rate of behavioral emotion recognition for dogs is also relatively low, and some research work still lacks real application scenarios.

Recently, deep learning models such as Xception and ResNet have begun to be used in pet expression emotion recognition. This paper proposes a pre-trained multi-modal transformer emotion detection system, which is first pre-trained on the human emotion detection dataset, which includes speech and facial expression data, and then uses the labeled animal voice and expression data as small Sample task data, this approach leverages unlabeled corpora for pre-training, satisfying the requirement to adequately train model parameters and prevent model overfitting, and then use the representations of these models for small-sample tasks.

The rest of this paper consists of: the part II introduces the multimodal-based emotion detection method, the part III proposes the pre-trained multi-modal transformer emotion detection model, and followed by the experimental results, finally is the conclusion.

2 MULTIMODAL-BASED EMOTION DETECTION MODEL

2.1 Emotion Detection

Affective computing began in 1995 and refers to the computing work related to emotions and emotions, mainly using automatic classification of emotion detectors, including: collecting information, extracting features and training models, and finally obtaining pattern recognition for emotions or emotions (Garcia-Garcia, Penichet, & Lozano, 2017). With the rapid expansion of Internet content, emotion detection for human expressions and speech can help improve the fluency and naturalness of human-computer interaction. In addition to the content in the language, it also includes voice prosody, voice intonation, body posture, gestures, etc., which can be expressed. The prosody can also be used for text accent or sentence segmentation, etc. It can also be distinguished in terms of fundamental frequency and intensity (Yu et al, 2001).

Different emotion models mainly come from the definitions of psychologists. Emotions such as happiness are adequately expressed in speech, but the expression of other emotions in speech may be inaccurate or insufficient. In the definition of psychology, the VAD space is used to define the emotional dimension representation. These three dimensions represent the degree of different emotions from one end to the other. This measure can be used to continuously represent emotions (Soleymani, 2015). An application scenario of emotion detection is to determine the emotional response of the audience when watching a movie. By using devices such as mobile phones to obtain the audience's voice and facial expressions while watching the movie, as well as using other physiological measurement devices, the audience's emotional response to the movie can be obtained. In this type of experiment, it is found that the individual's reflection is uncertain, and the collective emotion The response was also very inconsistent (Kim et al, 2021).

Fig.1 depicts the representation of emotions in several fields, which originated in psychology and philosophy, and were later explained using brain regions and neural mechanisms (Sailunaz et al, 2018). From a scientific point of view, human emotions can

be divided into groups, and the combination of these groups and parameters is defined as a model of emotion. Analysis of emotions from different sources has been widely studied by researchers. With the abundance of Internet resources, researchers have begun to use various information on social media for emotion analysis, such as voices and gestures and facial expressions and texts, as well as these vectors A combination of both is used to form multimodal data and perform emotion detection from them.

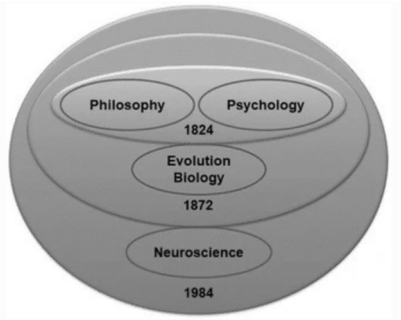

Figure 1: Evolution of emotion in various fields (Sailunaz et al, 2018).

2.2 Pet Emotion Detection

For the detection of pet emotions, some researchers use the method of pet group knowledge (Thibault, Bourgeois, & Hess, 2006). When domestic pets such as cats and dogs exhibit emotions similar to humans, humans can use expressions that describe human emotions, such as anger or happiness, to describe the pet's behaviour and respond to it.

Figure 2: Examples of cats showing interest (Thibault, Bourgeois, & Hess, 2006).

Some sensors, such as gyroscopes or accelerometers, exist in human life in an embedded way, in the form of wearable devices, which are also widely used in the field of animal detection and health management (Hussain et al, 2022). For example, the walking of horses and dogs can already be detected by wearable devices. These monitoring animal activities and behaviours and building animal behaviour patterns based on them belong to the research field of activity recognition, but because the normal activities of animals are different from humans, these can be Wearables also cannot be the same as human wearables. The researchers use gyroscopes and accelerometers, combined with CNN models to process these data from animal physical activity, and detect various activity categories of animals by extracting different characteristics of these data.

Pet-based emotion recognition generally focuses on cats and dogs (Cheng et al, 2022). For example, the cat's state can be classified according to the pattern in the cat's bark. The dog's bark can also be classified and recognized by extracting features and using CNN. In addition, the facial expressions of pets can also be used as the basis for emotion recognition. Changes in the nose or ears on the face are related to the pet's emotions, and under certain conditions, it can have a richer expression.

2.3 Multi-Modal Emotion Detection

Video, image, sound, and text are all carriers of information, and these carriers are traditionally handled separately by type. Recently, researchers have found that mixed processing of these carriers, such as multimodal systems, can more effectively obtain information carried by multiple carriers in the context, and good research results have been achieved in the fields of dialogue emotion detection, such as Multilogue-Net model (Shenoy & Sardana, 2020). This type of model can better improve the emotional cues expressed in the dialogue process by adding basic prior knowledge in the field and constructing some related data models such as similar intentions.

Recent multi-label sentiment detection research treats itself as a classification problem and introduces prior knowledge as added information (Ju et al, 2020), these methods may not be suitable for scenarios without prior knowledge. BERT and other sequence generation methods are widely used for multi-label sentiment detection, or to add more labels under a specific representation, and also increase the relationship between labels in multiple modalities on the relationship between labels.

In order to detect human emotional states, physiological measures such as EEG and other devices are combined with human information from audio and video for emotion recognition (Kim & Song, 2018). In previous human emotion recognition competitions in the wild, most machine learning techniques also required experts to select features,

while deep neural networks dominate today's human emotion recognition competitions in the wild.

The paper (Meng et al, 2022) proposes to decompose the video into visual part and audio part, and represent the video part with continuous images. In the emotion classification task, the facial expressions in these images are marked, and the audio parts corresponding to these images are also extracted. Out, these audio and visual information are multimodal context modelling, and the final part is the output of the fully connected layer. The video is first divided into

$$[n/p] + 1 \quad (1)$$

segments, where the i-th segment is defined as 1

$$\{F_{(i-1)*p+1}, \dots, F_{(i-1)*p+l}\} \quad (2)$$

Given visual features f_i^v and audio features f_i^a corresponding to the i-th segment, they are concatenated into the final layer to obtain multimodal features f_i^a. It can be expressed as:

$$\{F_{(i-1)*p+1}, \dots, F_{(i-1)*p+l}\} \quad (3)$$

where W_f and b_f are the tuning parameters.

3 PROPOSED PRE-TRAINED MULTI-MODAL TRANSFORMER EMOTION DETECTION MODEL

With the increasing attention of human beings to animal welfare, it has been widely recognized that some animals have emotions such as happiness (Mendl et al, 2009). Although these emotions are far from human emotions, the analysis of emotions in animals can help improve the level of animal welfare and help in the medical treatment of animals. For animals, emotions may be one of the factors affecting their survival and reproduction. Emotions can help animals choose their own behaviour and ensure maximum access to resources and avoidance of harm.

In the fields of psychology and cognition, the question of whether animals differ in intelligence and abilities such as learning or memory has been studied (Paul et al, 2020). Brain size and specific regions of the brain are both associated with cognitive ability, and there are many differences between the brains of different animals, making it difficult to adequately compare quantitative brains, whereas behavioural demands are more likely to predict these changes. Researchers believe that animals have the ability to choose resources, and this ability to choose comes from some human-like emotions, such as happiness, and the choice or abandonment of resources with the support of these emotions.

In conclusion, it is reasonable to perform individual or comprehensive emotion detection for multiple information of animals. This paper proposes a pre-trained multi-modal transformer emotion detection system, which is first pre-trained on a human emotion detection dataset including speech and facial expression data, and then takes the labelled animal voice and expression data as small-sample task data, This approach utilizes an unlabelled corpus for pre-training, satisfying the requirement to adequately train model parameters and prevent model overfitting, and then uses representations of these models for small-sample tasks.

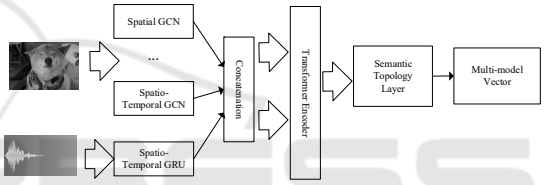

Figure 3: Proposed multi-modal transformer emotion detection system.

As shown in Fig.3, the model first pre-trains the multi-modal data of the human emotion detection dataset and extracts the multi-modal feature representation, and continues to train the small-sample task data on the multi-modal model. The layer semantic topology layer transfers the pre-trained model to the few-shot task. The semantic topology layer is mainly based on the topological connection, which is higher-level and abstract than the image semantic information, and forms a topological graph with emotional semantics by extracting the topological relationship between emotional expression elements.1

4 EXPERIMENTS AND RESULTS

This paper uses a dataset with pet expressions and audio from an online video platform, which contains a total of 50 videos with a total of about 110 minutes. The video is edited and the image size is normalized, each frame is marked manually, and some uncertain image data is deleted. A ratio of 8:2 is used to

distinguish the training set and the test set. The model proposed in this paper is compared with the single-modal ResNet, ConvNet, and CNN-Transformer in the experiments.

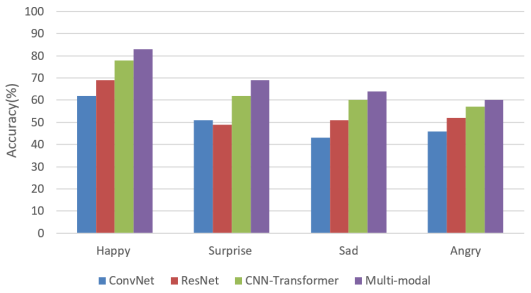

Figure 4: The average test accuracy of emotion detection.

The average test results of these methods on annotated datasets are shown in Figure 4. As can be seen from the figure, the Sad and Angry of ConvNet and ResNet are relatively low. The reason is that Sad and Angry are indistinguishable. The Happy of CNN-Transformer has obvious advantages over the former two, and there is still the problem that Sad and Angry are indistinguishable; In contrast, the accuracy rate of Sad and Angry of the multi-modal transformer structure proposed in this paper is slightly improved.

It can be considered that due to the fusion of multimodal information and the ability of the transformer to map to the same space to calculate the similarity, coupled with the abstraction ability of semantic topology, the discrimination ability of emotion classification in these four emotions is further improved.

5 CONCLUSION

The field of deep learning based on computer audio and video is developing rapidly. Combining language and visual information to carry out emotion detection and recognition is a new research hotspot in the field of artificial intelligence. In addition to detecting and recognizing human emotions, emotion recognition and management for pets has begun to receive attention. By collecting and processing pet sounds and corresponding facial expressions, it is possible to understand the pet's current condition and respond positively to these conditions. This paper proposes a pre-trained multi-modal transformer emotion detection system, which is first pre-trained on a human emotion detection dataset including speech and facial expression data, and then takes the labelled animal speech and expression data as small-sample task data, The method utilizes an unlabelled corpus for pre-training, which satisfies the requirements of adequately training model parameters and preventing model overfitting, and finally uses the representations of these models for few-shot tasks. Experimental results on video datasets show that the proposed multimodal Transformer structure has good accuracy compared to other algorithms.

REFERENCES

Bennett, V., Gourkow, N., & Mills, D. (2017). Facial correlates of emotional behaviour in the domestic cat (Felis catus). *Behavioural Processes. 141. 10.1016/j.beproc.2017.03.011.*

Cheng, W. K., Leong, W. C., Tan, J. S., Hong, Z. W., & Chen, Y. L. (2022). Affective Recommender System for Pet Social Network. *Sensors (Basel, Switzerland), 22(18), 6759. https://doi.org/10.3390/s22186759*

Garcia-Garcia, J., & Penichet, V., & Lozano, M. (2017). Emotion detection: a technology review. *1-8. 10.1145/3123818.3123852.*

Ho, J., Hussain, S., & Sparagano, O. (2021). Did the COVID-19 Pandemic Spark a Public Interest in Pet Adoption?. *Frontiers in veterinary science, 8, 647308. https://doi.org/10.3389/fvets.2021.647308*

Hussain, A., Ali S, S., Abdullah, M., & Kim, H. (2022). Activity Detection for the Wellbeing of Dogs Using Wearable Sensors Based on Deep Learning. *IEEE Access. 10. 10.1109/ACCESS.2022.3174813.*

Ju, X., Zhang, D., Li, J., & Zhou, G. (2020). Transformer-based Label Set Generation for Multi-modal Multi-label Emotion Detection. *Proceedings of the 28th ACM International Conference on Multimedia.*

Kim, D. H., & Song, B. C. (2018). Multi-modal Emotion Recognition using Semi-supervised Learning and Multiple Neural Networks in the Wild. *Journal of Broadcast Engineering, 23(3), 351–360. https://doi.org/10.5909/JBE.2018.23.3.351*

Kim, E.S., Bryant, D.G., Srikanth, D., & Howard, A.M. (2021). Age Bias in Emotion Detection: An Analysis of Facial Emotion Recognition Performance on Young, Middle-Aged, and Older Adults. *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society.*

Mendl, M.T., Burman, O.H., Parker, R.M., & Paul, E.S. (2009). Cognitive bias as an indicator of animal emotion and welfare: Emerging evidence and underlying mechanisms. *Applied Animal Behaviour Science, 118, 161-181.*

Meng, L., Liu, Y., Liu, X., Huang, Z., Jiang, W., Zhang, T., Deng, Y., Li, R., Wu, Y., Zhao, J., Qiao, F., Jin, Q., & Liu, C. (2022). *Multi-modal Emotion Estimation for in-the-wild Videos. ArXiv, abs/2203.13032.*

Paul, E. S., Sher, S., Tamietto, M., Winkielman, P., & Mendl, M. T. (2020). Towards a comparative science of emotion: Affect and consciousness in humans and

animals. *Neuroscience and biobehavioral reviews, 108, 749–770. 10.1016/j.neubiorev.2019.11.014*

Sailunaz, K., Dhaliwal, M., Rokne, J., & Alhajj, R. (2018). Emotion Detection from Text and Speech - A Survey. *Social Network Analysis and Mining (SNAM), Springer. 8. 10.1007/s13278-018-0505-2.*

Shenoy, A., & Sardana, A. (2020). Multilogue-Net: A Context Aware RNN for Multi-modal Emotion Detection and Sentiment Analysis in Conversation. *10.18653/v1/2020.challengehml-1.3.*

Shou, Q., Xu, Y., Jiang, J., Huang, M., & Xiao, Z. (2021). Detection of Basic Emotions from Cats' Meowing. *10.1007/978-981-16-1649-5_13.*

Singh, B., Dua, T., Sharma, D., & Adane, A. (2020). Animal Emotion Detection and Application. *Midas.*

Sinnott, R, O., Ackelin, U., Jia, Y., Sinnott, E, R, J., Sun, P., & 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). Brisbane, Australia, 2021, pp. 1-6, 10.1109/CSDE53843.2021.9718465.*

Soleymani, M., Asghair-Esfeden, S., Fu, Y., & Pantic, M. (2015). Analysis of EEG Signals and Facial Expressions for Continuous Emotion Detection. *IEEE Transactions on Affective Computing. 10.1109/TAFFC.2015.2436926.*

Thibault P., Bourgeois P., & Hess U. (2006). The effect of group-identification on emotion recognition: The case of cats and basketball players[J]. *Journal of Experimental Social Psychology, 2006, 42(5): 676-683. 10.1109/CSDE53843.2021.9718465.*

Yu, F., Chang, E., Xu, Y., & Shum, H. (2001). Emotion Detection from Speech to Enrich Multimedia Content. *550-557. 10.1007/3-540-45453-5_71.*