Author/s:
Sinnott, RO;Aickelin, U;Jia, Y;Sun, PY;Susanto, R

Title:
Run or Pat: Using Deep Learning to Classify the Species Type and Emotion of Pets

Date:
2021-01-01

Citation:
Sinnott, R. O., Aickelin, U., Jia, Y., 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, 00, EEE. https://doi.org/10.1109/CSDE53843.2021.9718465.

Run or Pat: Using Deep Learning to Classify the Species Type and Emotion of Pets

Richard O. Sinnott, Uwe Aickelin, Yunjie Jia, Elizabeth R.J. Sinnott, Pei-Yun Sun, Rio Susanto
School of Computing and Information Systems
The University of Melbourne, Melbourne, Australia
Contact: rsinnott@unimelb.edu.au

Abstract - **Deep learning has been applied in many contexts. In this paper we present a novel application area: to detect the species type and emotion of pets with focus on a diverse set of dog and cat collections comprising 52 dog and 23 cat species. Building on an extensive collection of labelled images with over 300 images per species type, we explore a range of deep learning models to develop a classifier for species type and their associated emotion. We outline the realization of the technical solution delivered through a mobile application (iPhone/Android) and present results based on feedback based on real world adoption and utilisation by the broader mobile application community.**

Keywords – *Deep learning, image recognition, pets, emotion.*

1 Introduction

Throughout history, pets have played a key role in peoples lives from hunting, protection through to offering emotional support and companionship. The adoption of pets has significantly increased in the Covid-19 era. However, for many individuals understanding their pets and their different emotions can be challenging. Dogs and cats can often exhibit explicit and clear signs of emotion, e.g., aggression through snarling and barking in the case of dogs or hissing and baring fangs in the case of cats, however this is not always as clear and obvious. In this paper, we explore how deep learning can be applied to classify the different types of species of pet and their emotional status.

The rest of this paper is structured as follows. In Section 2 we focus on related work in the deep lelarning space with emphasis on animal species recognition and face-based emotion detection/classification. In section 3, we discuss the data set that was curated to support this work. In section 4 we identify the methodology and design decisions for the mobile application that was developed. In section 5 we present the real-world results based on utilization of the mobile application by end users. Finally in section 6 we draw conclusions on the work as a whole and identify possible areas of future work.

2 Related Work

Understanding and classifying pet emotions has historically been based on intuitive judgment. There has been little work on how to sense pet emotions quantitively in a reliable and automated manner. Different species of animals may have diverse traits that can make the characteristic of their emotion difficult to assess. Yet such information is key in understanding pet behavior and the impact and consequences that this might have, e.g., dogs that may be about to snap/bite or cats about to scratch should be avoided, or pets that may have anxious expressions may need to go out to do their business.

The recent rise of artificial intelligence, machine learning and now deep learning gives rise to new opportunities to tackle this issue. Deep learning is now widely used by many researchers and companies alike in areas as diverse as computer vision, image recognition through to natural language processing. Compared to traditional methods, which hitherto required manual feature selection and adjustment, deep learning models utilise neural networks comprising several layers, where each layer can transform input to find more accurate and abstract features than possible through manual feature selection [1]. The most common demands of deep learning in the area of image processing are *object detection* and *classification*, e.g., is there a cat in the picture? (detection), and if so, what species of cat is in the picture? (classification).

There are several methods that have been applied for object detection in the deep learning domain including Faster Region Proposal Convolutional Neural Networks (Faster R-CNN) [2], Mask R-CNN [3], You Only Look Once (YOLO) [4] and Single Snapshot Detection (SSD) [5]. Faster R-CNN and Mask R-CNN grew out of R-CNN [6]. The R-CNN architecture proposes regions of interest (RoI) pooling layers of a given image that are applied for extracting fix-length feature vectors. Each output of the RoI layer is fed into fully connected layers as inputs that branch into two output vectors with a final (*softmax*) layer used for final

classification. However, region proposals can consume considerable time and computational resources. Mask R-CNN provides a framework for object instance segmentation where each RoI extends Faster R-CNN. It is used to predict object masks based on bounding boxes. Mask R-CNN not only shows the bounding box and class label, but also an object mask suitable for image recognition problems through addition of a quantization-free layer (*RoIAlign*). With this layer, the whole instance segmentation process can preserve correct spatial information up until the end result production thereby improving the accuracy of the mask location. The *RoIAlign* layer is designed to align extracted features with the inputs without the need for strict quantization. This quantization does not impact label classification, but it can lead to a misalignment between extracted features and the RoI. To avoid this, a new interpolation layer is used to replace the RoI boundaries quantization.

R-CNN-based models are based on two phases: region proposals and subsequent classification of the contents of those proposals. There are numerous single-phase approaches that have been put forwarded including YOLO and SSD. In these models each stage learns a feature map, and then carries out border regression and classification on the proposed map contents. There have been many approaches to reduce the size of the models to work in limited computational environments, e.g., IoT devices of mobile phones. MobileNet as one example provides a lightweight deep neural network model that is suitable for portable devices and embedded vision applications [7]. It achieves this by reducing the number of parameters required however this can impact on the overall accuracy [8].

The field of human emotion recognition has been explored extensively. Picard put forward the concept of Affective Computing through an approach based on recognizing, translating, processing, and mimicing human emotions [9]. The Emotion Recognition in the Wild (EmotiW) competition draws teams from around the world to classify and predict human emotions. While there are several categories each year, the category that has the most submissions are in the audio-video emotion recognition sub-category. In 2018, the best performing team in the audio-video category achieved an accuracy of 61.87% [10] and in 2019, the winning team achieved an accuracy of 63.39% [11].

Outside of the realm of deep learning, Kirana et al [12] adopted an approach based on the Viola-Jones Algorithm for emotion recognition. This algorithm provides a fast face detection algorithm based on passing parts of a facial image through a decision tree made up of a series of filters. If the provided part of the image makes it to the end of the decision tree, it is considered to be a face. Kirana used this approach for both facial detection and emotion recognition and achieved 76% accuracy. However, they only considered two emotional states: enthusiastic and bored.

Shojaeilangari et al [13] proposed an approach to facial emotion detection based on sparse learning where each face image was first turned in to a sparse representation of itself which was then input to a feedforward neural network. They achieved an accuracy on 66.5% on the EmotiW 2013 dataset.

In the area of animal emotion detection, there has been a limited number of published articles. Steagall et al [14] focused on identification of pain in cats, however this work was based on human classification and not based on machine learning. Belin et al analyzed cat and monkey calls and divided them into positive and negative emotions but the number of classifiers in this work was limited [15]. Molnar applied machine learning to analyze a range of different behaviors of dogs, but the overall recognition rate was low [16].

For species classification and emotion detection, deep learning offers numerous advantages for automated classification however this depends greatly on the data that is used to train the models.

3 Dataset Preparation

For deep learning tasks, the performance of the underlying neural network depends on the quantity and quality of the input data. For individual pet breed detection, a sufficiently pure dataset with correctly annotated labels is essential. Although many open-source dataset websites provide related dog breed images, most of them are incomplete and unsuitable for deep learning, e.g., they contain only a few images of specific breeds, or the breeds are labelled incorrectly or indeed ad hoc images have crept into the labelled data set, e.g., cartoon characters or paintings/T-shirts with specific species of animal are included in the datya set.

In this work we focus on establishing the data collection of individual dog and cat species types based on data from Google ImageNet [17] and the Stanford Dogs dataset [18]. Our goal was to create a data set comprised of as many species of dogs and cats as possible. Clearly it is impossible to cover every possible breed of dog and cat, especially since there are often mongrels / mixtures of species that might arise. Our focus was therefore on the most prevalent dogs and cats that people would most likely own. We identified 52 breeds of dog and 23 breeds of cat. These included dog species as diverse as Basset Hounds and Dachshunds to Siberian Huskies and cats as diverse as Abyssinian and Maine Coone to Sphynx.

As noted, the raw images for these dogs and cats were acquired from ImageNet and the Stanford Dog dataset. A key part of the work involved data normalisation, data cleaning and data labelling to establish a pure and feature rich data set. As noted, for the raw images, a large proportion of image data was invalid for many reasons. To address this, we filtered the raw images based on a range of practices. Firstly, the original scale (both the width and the height) of the raw images needed to be larger than a given size. Although it is possible to reshape the images, this process gives rise to a loss in the quality of the original image, which influences the classification performance. To avoid excessive disturbance of the image content, the percentage of useful content within an image must be greater than a fixed score, otherwise the model has a high probability to learn parameters for features that are useless and giving rise to erroneous classifications. As an example, Figure 1 shows an image of a Chiahuahua in the original image and the annotated bounding box associated with that image. As seen, the proportion of useful information, i.e., the actual dog face, forms a very small part of the original image. The annotated bounding box (shown in red) is used to identify the part of the image of interest. If the original image was used as input to the model, then there is a great chance that it would learn parameters related to the background as part of its classification. To avoid this, the images are clipped according to the bounding box information, i.e., the extracted image on the right-hand side is used for training the model.

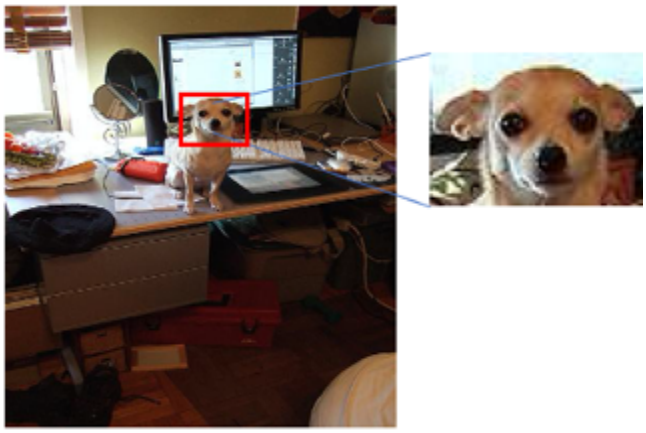

Figure 1: Original Image and Extracted Image

To increase the amount of data used for training and testing the model, a range of data augmentation techniques have been applied. These include rotating and flipping images, changing the aspect ratio and changing the contrast ratio. All input images were reshaped to a similar scale to ensure the convolutional layers were able to generate feature maps of the same size for any input images. As an example, Figure 2 shows how we reshape images whilst maintaining the stochasticity of the original data needed for the deep learning models.

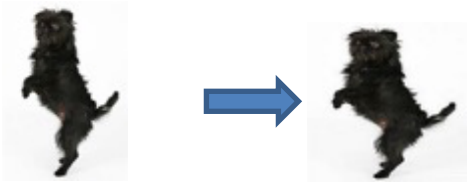

Figure 2: Reshaping images

One key aspect of the emotion detection is to gauge what emotion a given animal is exhibiting at the given time. In this work, we focused on a fixed set of emotion types: *happy, neutral, anxious, sad* and *unsettled*. The original prototype of the application had more emotion type classifications, e.g., angry, however the data sets that were utilised did not have an equal distribution of emotion types. Thus, the vast majority of dogs and cats in the original source data sets exhibited happy or neutral emotions. Furthermore, this classification was based solely on the face of the animal and not the whole body. It is often the case that animals exhibit their emotion using their whole bodies. Thus, dogs and cats can raise their hackles (*piloerection*) which causes the animal to appear larger by their fur and tails being raised. This is typically used as a visual warning to other animals (or people). The vast majority of dogs/cats in the data set did not have examples of such physical expressions however, hence the work focused entirely on the facial expressions.

After image preprocessing, there were at least 300 images per species of animal. We separated the set of images into a *training set*, a *development set*, and a *testing set* in the ratio 8:1:1. This is aligned with standard practice in deep learning training models.

4 Methodology and Mobile Application Development

The work explored several deep learning models including YOLO and Faster-RCNN. Originally it was planned to deliver the solution as a standalone mobile application, i.e., with no server-side support. However due to the computational restrictions imposed by mobile devices, the early models explored had limited accuracy. Instead, it was decided to deliver the application as a mobile application offering a lightweight client front end that would allow users to take pictures and send them to the server for actual classification. The server and associated back-end database and models were deployed on the University of Melbourne Research Cloud (https://dashboard.cloud.unimelb.edu.au).

This resource is freely available to all researchers at the University of Melbourne.

A range of experiments were conducted in applying different deep learning models for pet emotion detection [19]. This included exploration of models such as Xception, VGG16, ResNet-152, Inception-v3, InceptionResNetV2 and MobileNetV2 together with a detailed exploration of training parameters such as the batch size, loss function, kernel size, dropout and pooling layers and different fully connected layer options [19].

Eventually the finalized model was based on YOLO due to the accuracy and speed of classification. The mobile application itself was kept deliberately simple/intuitive. The mobile application has four main windows as shown in Figure 3. The end user can take a picture or select one from their gallery (Figure 3 left). Once taken/selected, the image is then sent to the Cloud-based server hosting the trained model for classification for results to be returned. As shown in Figure 3 (left middle), the results include both a bounding box around the pet's head, the type and species of pet along with the confidence level of the prediction as well as the prediction of the emotion and the confidence in the level of prediction.

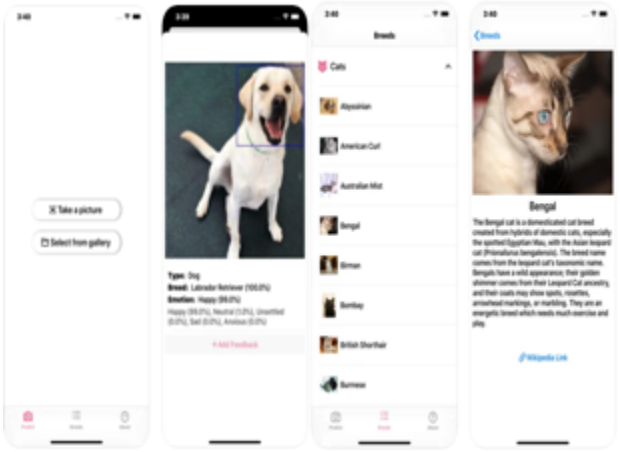

Figure 3: Mobile Application for HappyPets

The mobile app also allows for end users to provide feedback on the app itself. This includes whether the species prediction was correct as well as whether the emotion was correct (Figure 3 left middle bottom). The app also includes the catalogue of dogs and cats that the model has been specifically trained to classify (Figure 3 right middle/right). The app supports multiple pet classifications at the same time as shown in Figure 4.

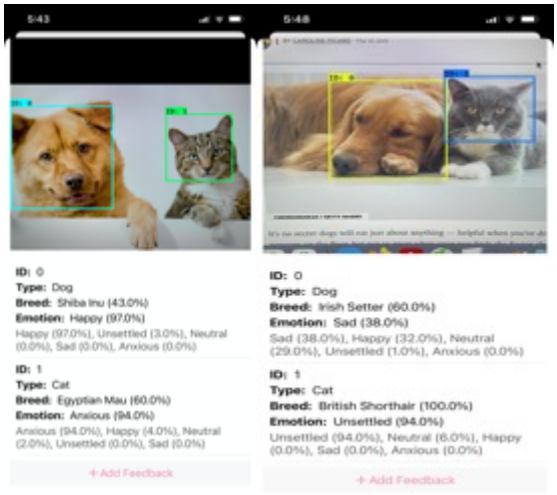

Figure 4: Multiple Pet Classification

5 Results

The mobile app has had considerable uptake by the broader community. It has been used for over 113k predictions with most end users taking pictures of their dogs as shown in Figure 5. This includes images with multiple pet faces in the image as well as other challenges, e.g., the faces of the owners or no pets at all.

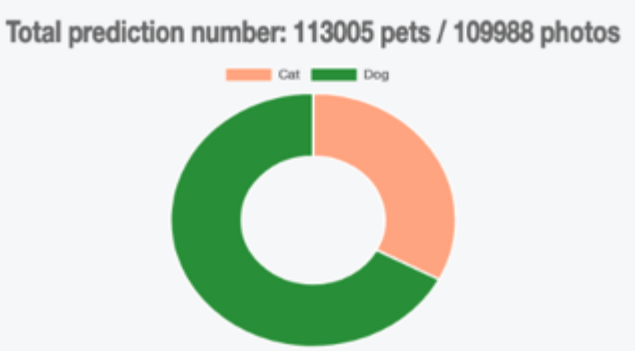

Figure 5: Utilisation of HappyPets

Not all people that use the app provide feedback on whether the model correctly identified the species and the emotion of the pet however. At present feedback on the accuracy of the models has been received 5166 times as shown in Figure 6.

Figure 6: Amount of Feedback for HappyPets

This feedback provides the real-world experience of how well the model performs. The results for classifying cat and dog species are shown in Figure 7 and Figure 8.

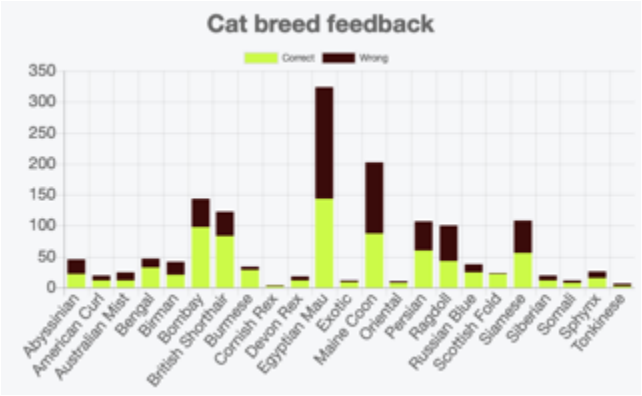

Figure 7: Accuracy of Cat Breed Classification

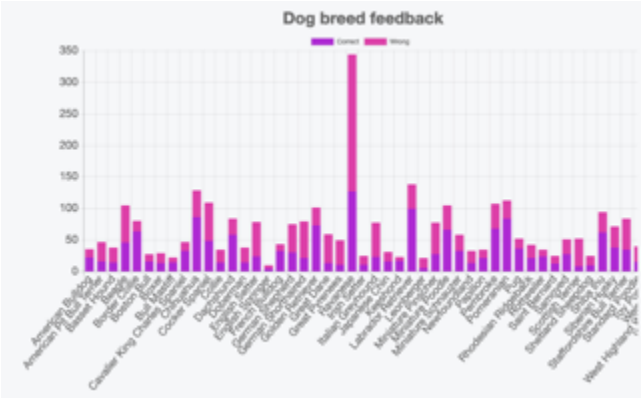

Figure 8: Accuracy of Dog Breed Classification

As seen, the overall accuracy of the dogs and cats varies. Often this is caused by the number of species of pets that have been chosen. Thus, the model has been trained to identify a specific number of species (52 dogs and 23 cats). Many users have attempted to use the app to identify species that are not supported. This includes pictures of non-pedigree dogs and cats. Nevertheless, the results are impressive since the accuracy is based on identifying the correct species amongst all of these that are supported and all other species that are not supported by the model.

The accuracy for the pet emotion is much better as shown in Figure 9 and Figure 10. As seen, the vast majority of cats and dogs are classified with the correct emotion. It can also be observed that the vast majority of predictions are based on classification of pets as being happy or neutral. This is primarily due to the vast majority of labelled data being associated with pets exhibiting happy/netural emotions.

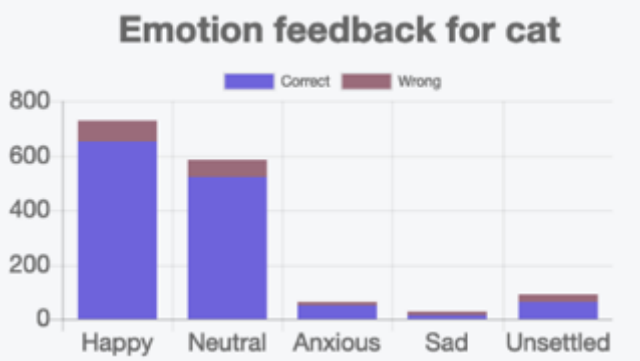

Figure 9: Accuracy of Cat Emotion Classification

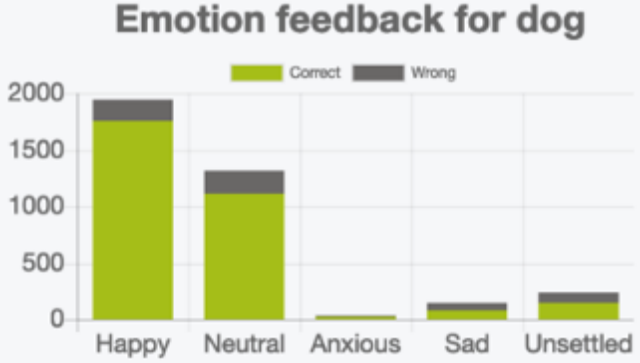

Figure 10: Accuracy of Dog Emotion Classification

6 Conclusions and Future Work

In this paper, we have presented a practical application of deep learning to classify the species type and emotion of an extensive collection of dog and cat species. The work has resulted in the production of a mobile application available in both the iPhone AppStore and Android Google Play. The overwhelming feedback on both applications has been positive with ratings of 4.7/5.0 on the Apple appStore.

There are many extensions and refinements to this work. Clearly extending the mobile applications to include more cat and dog species would be an obvious extension. Similarly, upgrading the model to leverage current state of the art models and the capabilities that they offer is another

extension. Thus, YOLOv5 now supports automated data augmentation through a dataloader API. This allows to extend the amount of data to maximise the subsequent precision based upon weightings associated with different data augmentation techniques as well as leveraging more refined approaches such as mosaicking and cropping and zooming.

There are many other emotions that pets can exhibit beyond thoese described here. This might include boredom or pain. Work is ongoing on the latter with automated (deep learning) based approaches for detecting cat pain. This is based on a collaboration with a major vet practice within Melbourne. The challenge is to capture sufficient data to train the models, however.

Extending the applications for pet owners specifically is another extension to the work. This might be used for keeping a pet diary for example of the day-to-day moods of pets by their owners.

The HappyPets mobile applications are available for download at https://apps.apple.com/au/app/happy-pets/id1515202735 (Apple iPhone) and https://play.google.com/store/apps/details?id=au.edu.unimelb.eresearch.happypets for the Android platform.

Acknowledgments

This research was undertaken using the LIEF HPC-GPGPU Facility hosted at the University of Melbourne. This Facility was establishe with the assistance of LIEF Grant LE170100200.

References

[1] Pak, M. and Kim, S., 2017, August. A review of deep learning in image recognition. 4th international conference on computer applications and information processing technology (CAIPT) (pp. 1-3). IEEE.

[2] Javier, R. Faster R-CNN: Down the rabbit hole of modern object detection. https://tryolabs.com/blog/2018/01/18/faster-r-cnn-down-the-rabbit-hole-of-modern-object-detection/.

[3] He, K., Gkioxari, G., Dollár, P., & Girshick, R. Mask R-CNN. In Proceedings of the IEEE international conference on computer vision, 2017 (pp. 2961-2969).

[4] Redmon, J., Divvala, S., Girshick, R., and Farhadi, A., 'You Only Look Once: Unified, Real-Time Object Detection', arXiv:1506.02640 [cs], Jun. 2015.

[5] Liu, W., et al., 'SSD: Single Shot MultiBox Detector', arXiv:1512.02325 [cs], vol. 9905, pp. 21–37, 2016.

[6] R. Girshick, R., Fast R-CNN. In Proceedings of the IEEE International Conference on Computer Vision (pp. 1440-1448).

[7] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L. C. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4510-4520).

[8] Koonce, B., 2021. MobileNetV3. In *Convolutional Neural Networks with Swift for Tensorflow* (pp. 125-144). Apress, Berkeley, CA.

[9] Picard, R.W., 2003. Affective computing: challenges. *International Journal of Human-Computer Studies*, *59*(1-2), pp.55-64.

[10] Dhall, A., Kaur, A., Goecke, R. and Gedeon, T., 2018, October. Emotiw 2018: Audio-video, student engagement and group-level affect prediction. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction* (pp. 653-656).

[11] Dhall, A., 2019, October. Emotiw 2019: Automatic emotion, engagement and cohesion prediction tasks. In *2019 International Conference on Multimodal Interaction* (pp. 546-550).

[12] Kirana, K.C., Wibawanto, S. and Herwanto, H.W., 2018, September. Facial emotion recognition based on viola-jones algorithm in the learning environment. In *2018 International seminar on application for technology of information and communication* (pp. 406-410). IEEE.

[13] Shojaeilangari, S., Yau, W.Y., Nandakumar, K., Li, J. and Teoh, E.K., 2015. Robust representation and recognition of facial emotions using extreme sparse learning. *IEEE Transactions on Image Processing*, *24*(7), pp.2140-2152.

[14] Evangelista, M.C., Watanabe, R., Leung, V.S., Monteiro, B.P., O'Toole, E., Pang, D.S. and Steagall, P.V., 2019. Facial expressions of pain in cats: the development and validation of a Feline Grimace Scale. *Scientific reports*, *9*(1), pp.1-11.

[15] Belin, P., Fecteau, S., Charest, I., Nicastro, N., Hauser, M.D. and Armony, J.L., 2008. Human cerebral response to animal affective vocalizations. *Proceedings of the Royal Society B: Biological Sciences*, *275*(1634), pp.473-481.

[16] Molnár, C., Pongrácz, P., Faragó, T., Dóka, A. and Miklósi, Á., 2009. Dogs discriminate between barks: The effect of context and identity of the caller. *Behavioural processes*, *82*(2), pp.198-201.

[17] Deng, J., Dong, W., Socher, R., Li, L.J., Li, K. and Fei-Fei, L., 2009, June. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition* (pp. 248-255). Ieee.

[18] Khosla, A., Jayadevaprakash, N., Yao, B. and Li, F.F., 2011, June. Novel dataset for fine-grained image categorization: Stanford dogs. In *Proc. CVPR Workshop on Fine-Grained Visual Categorization (FGVC)* (Vol. 2, No. 1).

[19] J. Zhang, X. Zhou, L. Li, D. Yu, Pet Smile, Masters Dissertation, University of Melbourne, 2020.