Multimodal recognition of dog social signals in human-animal interaction with pretrained model from a small-scale dataset

2310438 TRAN Anh Tuc

For centuries, dogs have been more than loyal companions. They have stood by humans in vital roles across military, security, therapy, and medical assistance. As our bond with dogs continues to grow, so does our responsibility to support not just their physical needs, but also their mental and emotional well-being. Understanding a dog's emotional state offers more than just insight into their inner world, it strengthens the human-dog relationship and leads to practical benefits. In parallel, advancements in AI have revolutionized the analysis and interpretation of unstructured data. Machine learning now enable the automation of a wide range of tasks and reveal vast, seemingly limitless potential. This thesis investigates machine learning approaches for classifying dog emotions based on vocal and behavioral patterns. This study focuses exclusively on YouTube videos where dog social signals are detected and labeled by animal experts. This thesis is conducted using a dataset that includes 7 classes of emotions per voice type.

This research field receives limited research attention, and there are few dog social signal datasets. In many real-world scenarios, such as bioacoustic analysis or emotion classification, labeled datasets are limited and expensive to obtain. Data scarcity is both a motivation and a challenge, particularly for data-driven machine learning. Research on dog emotions mainly focuses on unimodality, such as dog faces, poses, or voices. However, multimodal machine learning, leveraging different modalities that provide additional context, presents an effective method for emotion classification. Even though showing its remarkable ability, multimodal classification tasks often require large volumes of labeled data to effectively train deep learning models. Pretrained models show their powerful ability to present robust representations while addressing limited dataset problems with transfer learning. However, in unexplored research areas such as this one, there are few available in-domain pre-trained models. Despite advances in multimodal classification, current models struggle to effectively classify dog social signal due to limited datasets, background noise, and limited resources. This research aims to investigate effective methods for modeling dog social signals for classification on small-scale datasets using a multimodal machine learning approach. To achieve this goal, this study first develops a multimodal classification baseline for dog social signals using general pre-trained models as unimodal encoders. Additionally, a joint feature modulation (JFM) is proposed for multimodal fusion, which enables adaptively control the combination of pretrained representations, thereby preserving their rich information and even creating a

more informative fused representation, which in turn enhances overall results. Although pre-trained models provide a strong starting point, they often fail to fully capture the domain-specific characteristics of inputs. This leads to suboptimal performance, particularly in fine-grained classification tasks. To further improve the feature representations, this work proposes an approach to continue pretraining the unimodal encoder. The acoustic encoder is trained on available unlabeled data with initialization from cross-domain pretrained weights, enabling it to capture complex, discriminative acoustic features and make robust representations, thereby mitigating the problem caused by data scarcity and enhancing model robustness.

Multimodal model is built based on feature-level fusion with an unimodal encoder and nonlinear probing. AudioSet pretrained AST (Audio Spectrogram Transformer) is used to encode spectrogram input to acoustic embedding, and S3D pre-trained on the Howto100M dataset is used to encode video into visual embedding. Further ablation studies are conducted using human speech and image-pretrained models to evaluate the effectiveness of cross-domain transfer learning in improving the audio unimodal feature representation. Data augmentation and PCA visualization will assist in generalization and interpretability. After training, models will be evaluated using accuracy and F1-score.

The expected outcome is this multimodal-based classification model with higher accuracy and robustness compared to the majority, non-parametric handcrafted feature model, and unimodal baselines. Results confirm the effectiveness of this proposed multimodal pretrained model based approach. Moreover, compared to simple early fusion methods, the proposed JFM fusion demonstrates improved performance across both metrics, proving to be more effective in our setting. However, this has not met our expectations. To enhance the acoustic encoder and ultimately improve multimodal performance, this work explores continued pretraining to inject in-domain discriminative features into unimodal embeddings based on cross-domain pretrained models. Although this experiment remains incomplete due to various limitations, it demonstrates strong potential for future development.

In conclusion, these results, achieved with a small and noisy dataset under simple settings and a transfer learning framework, nonetheless highlight the potential of multimodal classification of dog social signals and serve as the baseline for future references on this task. This work aims to contribute a reproducible research pipeline for dog social signal prediction to contribute to human-computer-animal research, hopefully guiding the development of more robust and efficient approaches.