Interpreting Multimodal Machine Learning Models Trained for Emotion Recognition to Address Robustness and Privacy Concerns

Mimansa Jaiswal
University of Michigan
mimansa@umich.edu

Abstract

Many mobile applications and virtual conversational agents now aim to recognize and adapt to emotions. These predicted emotions are used in variety of downstream applications: (a) generating more human like dialogues, (b) predicting mental health issues, and (c) hate speech detection and intervention. To enable this, data are transmitted from users' devices and stored on central servers. These data are then processed further, either annotated or used as inputs for training a model for a specific task. Yet, these data contain sensitive information that could be used by mobile applications without user's consent or, maliciously, by an eavesdropping adversary. My work focuses on two major issues that are faced while training emotion recognition algorithms: (a) privacy of the generated representations and, (b) explaining and ensuring that the predictions are robust to various situations. Tackling these issues would lead to emotion based algorithms that are deployable and helpful at a larger scale, thus enabling more human like experience when interacting with AI.

1 Introduction

Virtual conversational agents strive to emulate human-like interaction to have more naturally flowing conversation (Metcalf et al. 2019). These agents often employ models that classify aspects of communication, including the classification of the emotional content of speech. The resulting predictions can then be used to bias response generation. Emotion classification is also used in mobile and web applications to identify heightened risk of suicidal ideation or mood fluctuations (Khorram et al.), for the purpose of tracking or intervention. Data are sent from users' devices, including mobile applications and Alexa or Google home devices (Piersol and Beddingfield 2019), and are stored on central servers for analysis. However, data transmitted from users' devices are vulnerable to data hacking and re-identification (Barbaro, Zeller, and Hansell 2006). A way to counter this issue in data collected by mobile or smart home applications is to generate a data representation on the device and then to transfer that representation to the server for additional processing. The benefit is that these representations can increase privacy by partially obfuscating the actual content of the conversation (Bengio, Courville, and Vincent 2013) but they still contain sensitive demographic information. These collected representations are also not indicative of the whole population. Various psychological factors affect how individuals express emotions. Yet, when we collect data intended for use in building emotion recognition systems, we often try to do so by creating paradigms that are designed just with a focus on eliciting emotional behavior. Algorithms trained with these types of data are unlikely to function outside of controlled environments because our emotions naturally change as a function of these other factors. These algorithms also rely on data annotated with high quality labels. However, emotion expression and perception are inherently subjective. As a result, annotations are colored by the manner in which they were collected. This process of ending up at a final emotion recognition model summarizes the three main issues with at scale usage: (a) Privacy of the data collected, (b) Variability in the data collected, and (c) Variability in perception of emotion. Unlike other machine learning algorithms that deal with objective output, such as speech to text or object recognition, emotion production and perception changes is influenced by the person, environment, and various other factors. This results in dealing with subjective outputs for optimizing machine learning algorithms, which not only makes it harder to process, but also, harder to evaluate in terms of performance. In my thesis, I aim to study some methods to counteract these pitfalls.

2 Related Work

The related work can be hence be divided into three streams. Recent research has examined privacy preservation in the context of neural networks. These efforts have primarily focused on ensuring that the input data are not memorized and cannot be retrieved given a deployed model, either by accounting for unintended memorization or adding random noise to either the aggregated dataset or to individual datapoints (Carlini et al. 2019). Previous work also looked at task-specific privacy preservation for a particular attribute in a dataset (Elazar and Goldberg 2018) and (Coavoux, Narayan, and Cohen 2018) for textual data, using minmax modeling, declustering, or adversarial training. On the other hand has looked at removing confounding factors as a graph

problem, with methods such as graph pruning (Molchanov et al. 2016), surgery estimation (Subbaswamy, Schulam, and Saria 2018) and counterfactual adjustments. In the context of neural networks, controlling for confounding factors during training is commonly achieved via the adversarial training paradigm (Ganin et al. 2016). The effective use of crowdsourcing for collecting reliable emotion labels has been an active research topic. (Burmania, Abdelwahab, and Busso 2016) investigated the trade-off between the number of annotators and underlying reliability of the annotations. Other work has looked at quality-control techniques to improve the reliability of annotations (Soleymani and Larson 2010).

3 Completed Work

In our paper (Jaiswal et al. 2019) at ICASSP '19 where we tackled variations occurring due to method of annotation, we conducted crowdsourcing experiments to investigate this impact of collection on both the annotations themselves and on the performance of these algorithms. We focus on one critical question: the effect of context. We present a new emotion dataset, Multimodal Stressed Emotion (MuSE), and annotate the dataset using two conditions: randomized, in which annotators are presented with clips in random order, and contextualized, in which annotators are presented with clips in order. The paper (Jaiswal, Aldeneh, and Mower Provost 2019) at ICMI '19 studied how the multimodal expressions of emotion change when an individual is under varying levels of stress. We hypothesize that stress produces modulations that can hide the true underlying emotions of individuals and that we can make emotion recognition algorithms more generalizable by controlling for variations in stress. To this end, we use adversarial networks to decorrelate stress modulations from emotion representations. We study how stress alters acoustic and lexical emotional predictions, paying special attention to how modulations due to stress affect the transferability of learned emotion recognition models across domains. In the paper at AAAI (Jaiswal and Provost 2019), we show how multimodal representations trained for a primary task, here emotion recognition, can unintentionally leak demographic information, which could override a selected opt-out option by the user. We analyze how this leakage differs in representations obtained from textual, acoustic, and multimodal data. We use an adversarial learning paradigm to unlearn the private information present in a representation and investigate the effect of varying the strength of the adversarial component on the primary task and on the privacy metric.

4 Future Work

I am primarily interested in (a) interpreting model behavior, (b) analysing how perturbations to input correlates with model outcomes, and, (c) using the previous two concepts to evaluate and define model trust or confidence score. I aim to understand how various emotion systems behave in the presence of varying linguistic and environmental contexts, and, how we can explain and ensure that the chosen response or action of an agent is correct and trustworthy, in terms of both, utility and privacy.

References

Barbaro, M.; Zeller, T.; and Hansell, S. 2006. A face is exposed for aol searcher no. 4417749. *New York Times*.

Bengio, Y.; Courville, A.; and Vincent, P. 2013. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*.

Burmania, A.; Abdelwahab, M.; and Busso, C. 2016. Trade-off between quality and quantity of emotional annotations to characterize expressive behaviors. In *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, 5190–5194. IEEE.

Carlini, N.; Liu, C.; Erlingsson, Ú.; Kos, J.; and Song, D. 2019. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th {USENIX} Security Symposium)*.

Coavoux, M.; Narayan, S.; and Cohen, S. B. 2018. Privacy-preserving neural representations of text. *arXiv preprint arXiv:1808.09408*.

Elazar, Y., and Goldberg, Y. 2018. Adversarial removal of demographic attributes from text data. *arXiv preprint arXiv:1808.06640*.

Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; Marchand, M.; and Lempitsky, V. 2016. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research* 17(1):2096–2030.

Jaiswal, M.; Aldeneh, Z.; and Mower Provost, E. 2019. Controlling for confounders in multimodal emotion classification via adversarial learning. *arXiv preprint arXiv:1908.08979*.

Jaiswal, M., and Provost, E. M. 2019. Privacy enhanced multimodal neural representations for emotion recognition. *arXiv preprint arXiv:1910.13212*.

Jaiswal, M.; Aldeneh, Z.; Bara, C.-P.; Luo, Y.; Burzo, M.; Mihalcea, R.; and Mower Provost, E. 2019. Muse-ing on the impact of utterance ordering on crowdsourced emotion annotations. In *2019 ICASSP*. IEEE.

Khorram, S.; Jaiswal, M.; Gideon, J.; McInnis, M.; and Mower Provost, E. The priori emotion dataset: Linking mood to emotion detected in-the-wild. In *Interspeech 2018*.

Metcalf, K.; Theobald, B.-J.; Weinberg, G.; Lee, R.; Jonsson, I.-M.; Webb, R.; and Apostoloff, N. 2019. Mirroring to build trust in digital assistants. *arXiv preprint arXiv:1904.01664*.

Molchanov, P.; Tyree, S.; Karras, T.; Aila, T.; and Kautz, J. 2016. Pruning convolutional neural networks for resource efficient transfer learning. *arXiv:1611.06440* 3.

Piersol, K. W., and Beddingfield, G. 2019. Pre-wakeword speech processing. US Patent App. 14/672,277.

Soleymani, M., and Larson, M. 2010. Crowdsourcing for affective annotation of video: Development of a viewer-reported boredom corpus. *SIGIR-Workshops*.

Subbaswamy, A.; Schulam, P.; and Saria, S. 2018. Learning predictive models that transport. *arXiv preprint arXiv:1812.04597*.