Privacy Enhanced Multimodal Neural Representations for Emotion Recognition

Mimansa Jaiswal, Emily Mower Provost
University of Michigan
{mimansa, emilykmp}@umich.edu

Abstract

Many mobile applications and virtual conversational agents now aim to recognize and adapt to emotions. To enable this, data are transmitted from users' devices and stored on central servers. Yet, these data contain sensitive information that could be used by mobile applications without user's consent or, maliciously, by an eavesdropping adversary. In this work, 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, defined here as the inability of an attacker to predict specific demographic information. We evaluate this paradigm on multiple datasets and show that we can improve the privacy metric while not significantly impacting the performance on the primary task. To the best of our knowledge, this is the first work to analyze how the privacy metric differs across modalities and how multiple privacy concerns can be tackled while still maintaining performance on emotion recognition.

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. (Huang et al. 2018). 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. ; Matton, McInnis, and Provost 2019; Gideon et al. 2019), for the purpose of tracking or intervention. Data are sent from users' devices, including mobile applications (Khorram et al.) 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). Eavesdroppers can use these data for identification of an individual and to gain access to sensitive information. 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). However, they still contain sensitive demographic information.

The implications of sensitive information leakage is profound: research has shown that discrimination occurs across variables of age, race, and gender in hiring, policing and credit ratings (Hajian and Domingo-Ferrer). (Abadi et al.) showed how adding random noise to aggregated dataset or individual samples can ensure defense against privacy attacks. But, previous research has shown that privacy induced using additional noise can often be exploited if the adversary has access to the network used to generate anonymity (Kifer and Machanavajjhala 2011). Therefore to ensure robustness, we consider a scenario of the attacker having access to the same embedding sub-network to generate the representations that will be used to train its attack network.

In this work, we focus on privacy in the context of emotion recognition. Emotion recognition provides an important test case because emotion production varies significantly across gender and race. As a result, the outputs of emotion recognition models are often highly correlated with these secondary demographic signals (Chaplin 2015; Soto and Levenson 2009). We design approaches to first measure leakage and then to counteract this leakage. We measure privacy in two ways: 1) using a privacy metric, which we define as the incapability of an attacker to recover demographic information from representations, and 2) by determining an adversary's ability to perform membership identification (Li et al.), defined as the ability to determine if a given user was in a dataset from which the emotion recognition models were trained (this can be harmful if the training data are collected in a sensitive context, such as counselling or therapy). We ask the following seven questions:

1. Does demographic leakage differ in umimodal and multimodal emotion recognition models?
2. How does the privacy metric change when a network is trained to preserve privacy?
3. How does emotion recognition performance change when networks are trained to preserve privacy?
4. How does the adversarial component's strength impact emotion recognition performance and the privacy metric?
5. Focusing on gender, how does the performance of emotion recognition change when a network is trained to preserve privacy?
6. Does the location of the adversarial component within a network affect the privacy metric and emotion recognition performance?
7. Does the privacy preserving paradigm help defend against other attacks such as membership identification?

Our results show that representations obtained for emotion recognition can be exploited by an adversary to predict sensitive variables given unimodal information (either audio or lexical). We further show that multimodal models contain even more sensitive information, as lexical and audio each encode different aspects of demographic information. We show that we can increase the defense against this attack by adversarially training representations to be invariant to gender. The novelty of this work is two fold: (1) we analyze how the demographic privacy of a representation differs across modalities and how it can be increased using adversarial paradigm; and, (2) we obtain privacy enhanced representations that defend against multiple privacy attacks while still maintaining performance on emotion recognition.

2 Related Work

Previous research has investigated methods to improve privacy in data collection. Early work focused on rule-based systems, which would identify patterns in text and replace them with random word tokens (Gomez-Hidalgo et al. 2010). Other methods included the addition of background noise or randomizing the order of sentences (Evfimievski 2002). These systems, though easy to interpret, are harder to scale to larger or varying distributions of datasets for they might necessitate an increase in the number of rules required and require expert input.

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. (Carlini et al. 2019) showed attackers could extract unique and secret sequences such as credit card numbers given models that are trained without accounting for unintended memorization. (Abadi et al.) proposed adding random noise to either the aggregated dataset or to individual datapoints to defend against membership query attacks. This method though, is usually either used for structured data or images and often incurs a cost in terms of a reduction in task performance.

Another line of work concentrates on fair algorithmic representation. Though the end goal isn't privacy, the methodology is similar. The aim is to create networks that are invariant to particular attributes, usually demographic information to obtain debiased word embeddings (Bolukbasi et al. 2016), ensure fairness pairities (Corbett-Davies and Goel 2018), and train fair hate speech classification (Davidson, Bhattacharya, and Weber 2019).

Previous research has looked at task-specific privacy preservation for a particular attribute in a dataset. For example, (Elazar and Goldberg 2018) investigated text-based privacy preservation for sentiment. (Zhao et al. 2019) looked at minmax modelling of the utility-privacy tradeoff by classifying gender as a primary task, while masking ethnicity and age. (Coavoux, Narayan, and Cohen 2018) looked into modelling privacy by declustering representations that fall under the same sensitive attribute subgroup.

Given most of the previous work on privacy preserving representations concentrates on just lexical information, we tackle the questions that arise from desiring privacy preservation in multimodal representations for emotion recognition. While the primary goal of most previous works has been to avoid unintentional inference by the application itself, we concentrate on minimizing the potency of an attacker to deliberately recover sensitive attributes from an invariant representation.

3 Datasets and Features

3.1 Datasets

We use four common emotion recognition datasets: MSP-Improv (Busso et al. 2017), MSP-Podcast (Lotfian and Busso 2017), Interactive Emotional Dyadic MOtion Capture (IEMOCAP) dataset (Busso et al. 2008), and Multimodal Stressed Emotion (MuSE) dataset (Jaiswal et al. 2019).

MuSE. The MuSE dataset was collected to understand the interplay between stress and emotion in natural spoken communication. It contains audio, video, thermal, physiological data and associated manual transcriptions. The dataset consists of 55 recordings from 28 participants, for a total of 2,648 utterances. For these recordings, emotion in the participant was induced via emotionally evocative monologue topics (Aron et al. 1997). Data selection was performed to reduce the dataset to include only utterances of length $[3, 25]$, inline with previous emotion datasets (Khorram et al.).

IEMOCAP. The IEMOCAP dataset was created to explore the relationship between emotion, gestures, and speech. Pairs of actors, one male and one female (five males and five females in total), were recorded over five sessions (either scripted or improvised) The data were segmented by speaker turn, resulting in a total of 10,039 utterances (5,255 scripted turns and 4,784 improvised turns). It contains audio, video, and associated manual transcriptions.

MSP-Improv. The MSP-Improv dataset was collected to capture naturalistic emotions from improvised scenarios while partially controlling for lexical content. The data of 8,438 sentences were divided into 652 target sentences, 4,381 improvised turns (the remainder of the improvised scenario, excluding the target sentence), 2,785 natural interactions (interactions between recordings of the scenarios), and 620 read sentences (emotional readings of the target sentences). It contains audio, video, and transcriptions derived from automatic speech recognition (ASR).

MSP-Podcast. The MSP-Podcast dataset was collected to build a naturlisitic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings. This was done using machine learning algorithms, which along with a cost-effective annotation process using crowdsourcing, led to a vast and balanced dataset. We use a pre-split part of the dataset which has been identified for gender of the speakers which comprises of 13,555 utterances. The dataset as a whole contains audio recordings.

3.2 Labels

Emotion Labels. Utterances in each of the four datasets were labeled using the dimensional descriptions of activation (calm vs. excited) and valence (positive vs. negative). Each utterance in the MuSE dataset was labeled on a nine-point Likert scale by eight crowd-sourced annotators (Jaiswal et al. 2019), who observed the data in random order across subjects. We average the annotations to obtain a mean score for each utterance, and then bin the mean score into one of three classes, defined as, {"*low*": [min, 4.5], "*mid*": (4.5, 5.5], "*high*": (5.5, max]} valence and activation. Utterances in IEMOCAP and MSP-Improv were annotated on a five-point Likert scale. The activation and valence values for were averaged over all annotations for an utterance and binned as: {"*low*": [1, 2.75], "*mid*": (2.75, 3.25], "*high*": (3.25, max]}. The labels for MSP-Podcast were annotated on a seven point Likert scale and averaged over all annotations for an utterance and binned as: {"*low*": [1, 3.75], "*mid*": (3.75, 4.25], "*high*": (4.25, max]}.

3.3 Features

Acoustic. We use Mel Filterbank (MFB) features, which are frequently used in speech processing applications, including speech recognition, speaker recognition, and emotion recognition. We extract the 40-dimensional MFB features using a 25-millisecond Hamming window with a step-size of 10-milliseconds. Each utterance is represented as a 40-dimensional time series. We z-normalize the obtained acoustic features by speaker.

Lexical. We use the word2vec representation (Mikolov et al. 2013) based on the transcriptions for MuSE and IEMOCAP, which has shown success in sentiment and emotion analysis tasks. We do not use MSP-Improv due to errors in ASR transcription or MSP-Podcast due to the lack of transcripts. We represent each word in the input as a 300-dimensional vector using a pre-trained word2vec model, replacing out-of-vocab words with the $\langle unk \rangle$ token. Each utterance is represented as a sequence of 300-dimensional feature vectors.

4 Experimental Setup

In this section, we describe the network architecture, the training recipe, and the metrics in consideration.

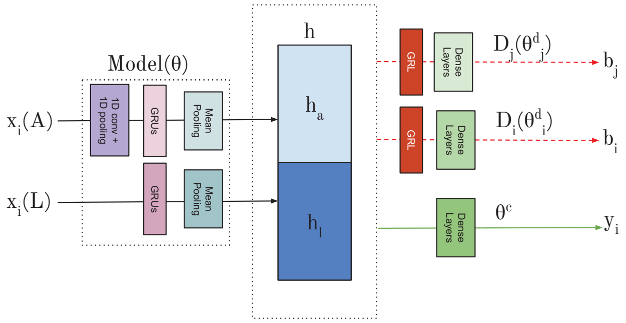

Figure 1: Privacy preserving network architecture.

4.1 Architecture

The objective of this system is to maximize the performance of the emotion classifier while minimizing the performance of the gender classifier (see Figure 1). The main network consists of three components: (1) embedding sub-network, $Model(\theta)$; (2) emotion classifier, θ^c and output y_i; and (3) gender classifier, $D_i(\theta_i^d)$, with output, b_i. We then disucss how an attacker network could maliciously use this information to obtain sensitive demographic information.

Main Network. The *embedding sub-network* uses a state-of-the-art multimodal approach in emotion recognition (Aldeneh et al. 2017) in which the acoustic and lexical information are processed separately and then joined after the application of modality-specific global mean pooling. The acoustic input stream $(x_i(a))$, where i represents an input frame (40-dimensional) and a represents acoustic, is processed using a set of convolution layers and Gated Recurrent Units (GRU), which are fed through a mean pooling layer to produce an acoustic representation (h_a). The lexical input $(x_i(l))$, where i represents an input word (300-dimensional) and l represents lexical, is passed through GRUs and pooled to obtain a lexical representation (h_l). For the multimodal setup, these two representations, (h_a) and (h_l), are concatenated (h). The representations (h, h_a, h_l) are of fixed-length given acoustic and lexical inputs. The *emotion classifier* takes in the representation $(h, h_a,$ or $h_l)$ as input and estimates valence or activation using a set of dense layers. The *gender classifier* estimates the gender label (i.e., male or female) using a set of dense layers.

Gender-Leakage. The main network is trained to unlearn gender. To achieve this goal, we use a Gradient Reversal Layer (GRL) (Ganin and Lempitsky 2014). GRLs are a common multi-task approach to train networks such that they are invariant to specific properties (Meng et al. ; Jaiswal, Aldeneh, and Mower Provost 2019). During the backward pass of the training phase, the GRL multiplies the backpropogated gradients by $-\lambda$ (i.e., the strength of the adversarial component). During the forward pass, the GRL acts as an identity function. To make the network invairant to gender, we place the GRL function between the embedding sub-network and the gender classifier. We obtain gender-invariant representations using the following loss function:

$$\widehat{\theta} = \underset{\theta_M}{min} \underset{\{\theta_{D^i}\}_{i=1}^N}{max} \chi(\widehat{y}(x;\theta_M), y) - \sum_{i=1}^{N} (\lambda_i . \chi(\widehat{b}(x;\theta_{D^i}), b_i))$$

where N is the number of targeted sensitive variables (here N=1). The loss function ensures that while the output components are trained to be good predictors, the representation is trained to be maximally good for the primary task (emotion) and maximally poor for the secondary task (gender).

Attacker Network. We assume that the attacker has access to a held-out dataset (either a different dataset or a section of the original dataset) with known gender labels. The attacker generates representations for this dataset using the previously described embedding sub-network. The network then learns to predict gender labels from the generated representations using a set of dense layers. Since the parameters used to construct the representation are fixed, the attacker only acts upon its own parameters to optimize the gender prediction linear loss. The purpose of the attacker's network is to recover gender information from representations whose labels are unknown. Though testing using a singular network isn't a guarantee of robustness of the representation to privacy attacks, for the scope of this paper, we use a feed forward network, one of the powerful learning methods on a fixed size static representation.

Model Variations. We use 12 variants of the network shown in Figure 1. We train combinations of the following setups: {*general-classification-model (Gen), privacy-preserving-classification-model (Priv)*} × {*activation, valence*} × {*uni-lexical, uni-acoustic, multimodal*}.

The general classification setup makes use of the embedding sub-network with text, speech, or both as input streams and the emotion classifier. The privacy preserving classification setup adds the gender classifier to the general setup.

Training. We implement models using the Keras library. (Chollet 2015). We use a weighted cross-entropy loss function for each task and learn the model parameters using the RMSProp optimizer. (Tieleman and Hinton 2012). We train our networks for a maximum of 50 epochs and monitor the validation loss for the emotion classifier after each epoch, stopping the training if the validation loss does not improve for five consecutive epochs. Once the training ends, we revert the network's weights to those that achieved the lowest validation loss on the emotion classification task. For the privacy preserving classification model, we ensure that the chosen model yields a chance unweighted average recall (UAR) for the gender classification task on the validation set. Finally, we train each setup three times with different random seeds and average the predictions over these runs to reduce variations due to random initialization.

We use validation samples for hyper-parameter selection and early stopping. The hyper-parameters that we use for the main network include: number of convolutional layers {3, 4}, width of the convolutional layers {2, 3}, number of convolutional kernels {32, 64, 128}, number of GRU layers {2, 3}, GRU layers width {32}, number of dense layers {1, 2}, dense layers width {32, 64}, GRL λ {0.3, 0.5, 0.75, 1}. For the adversarial emotion classification setups, we use the hyper-parameters that maximize the validation emotion classification performance while minimizing the validation gender classification performance. For the attacker's model, we use the following hyper-parameters: number of dense layers {2, 3, 4}, dense layers width {32, 64}. We report the UAR performance of our models, given the imbalanced nature of our data. (Rosenberg 2012).

4.2 Metrics

Performance. We define performance for emotion recognition as the ability of the model to correctly classify either activation or valence into 3 categories: low, medium, and high. We measure performance using UAR (chance is 0.33).

Demographic Leakage. Leakage is defined as the ability of a trained gender classifier to predict gender from the representations which are obtained when the network is simultaneously trained to perform the primary task.

Demographic Privacy Metric. We define the privacy metric as the inability of an attacker to be able to recover gender from the representations trained on a primary task. To test this, we use four phases of training.

1. We train the main network on a dataset (D1), represented by the pair (x_{D1}, y_{D1}), where x is the data input while y is the gender label. We obtain representations for this dataset ($h(x_{D1})$).
2. We consider that the attacker has access to another dataset or unused subset of the same dataset (D2) represented by the pair (x_{D2}, y_{D2}). We generate representations $h(x_{D2})$ for the pairs in this dataset using the embedding sub-network of the main network.
3. We train a model (M_{att}) to predict gender labels using the representations obtained in step 2, represented as $M_{att}((h(x_{D2}), y_{D2}))$.
4. Using the model obtained previously (M_{att}), we choose $h(x_{D1})$ as inputs, and measure the gender prediction capability of the attacker $UAR(M_{att}((h(x_{D1}), y_{D1}))$. The Demographic Privacy Metric of an attacker is then quantitatively defined as $1 - UAR(M_{att})$.

The range of the privacy metric goes from 0 (the attacker is always correct) to 0.5 (the attacker has a chance UAR).

Membership Identification. Membership identification is the possibility of an attacker being able to recognize if a speaker belongs to the training set. We assume that the adversary can obtain samples from a speaker from the same distribution as that for the training set. Consider that the adversary knows some speakers for whom representations definitely exist in the training set and some for whom they definitely don't. We test the possibility of membership identification using four steps:

1. We simulate the above using cross-validation. Given five speaker independent folds, we use three for the training set. From the remaining two folds, we add some samples of the selected speakers to the training set.

2. We consider that the attacker knows both, the speakers selected and not selected for training from set four $(s4)$, but has no information about this split for set five $(s5)$. The objective of the attacker is to predict whether speakers were selected for inclusion in the training set from $(s5)$.
3. The attacker trains a binary classification model comprised of dense layers (M_{att-mi}) using the representations obtained from dataset D1 as $M_{att-mi}(h(x_{D1}), `Yes')$. It obtains representations of the samples not used in training for the the selected speakers included in the training set and trains its model as $M_{att-mi}(h(x_{s4_selected}), `Yes')$ and for the speakers not included in training as $M_{att-mi}(h(x_{s4_selected}), `No')$. A speaker is saved from each label for validation.
4. We then define the UAR of the performance of $M_{att-mi}(h(x_{s5}))$ as membership identification.

5 Analysis

In all the tables, **U** is the unweighted average recall (UAR), and **U(M)** and **U(F)** represent the performance of the model for emotion recognition when gender is male and female respectively. Leakage in the model is represented by **L**, the lower the better, where chance leakage is 0.5. Privacy metric, represented by **P**, ranges from $[0, 0.5]$, and is the incapability of an attacker to obtain demographic information from the representation, the higher the better. Membership identification represented by **MI**, ranges from $[0.5$ (chance UAR),1], and is the capability of an attacker to identify if the subject belongs in the training set, for which the lower the value, the better. We code identify the datasets as follows: **Imp**-MSP-Improv; **Pod**-MSP-Podcast; **Iem**-IEMOCAP; and **MuS**-MuSE. All significance tests are paired t-test, with significance established (shown in bold) when **Benjamini-Hochberg adjusted** (FDR $= 5\%$) p-value< 0.05.

5.1 Question 1

Q: *Does demographic leakage differ in umimodal and multimodal emotion recognition models?*

HYPOTHESIS: *Multimodal representations leak more gender information than unimodal representations.*

Previous research has shown that different modalities have varying capabilities of capturing demographic information, such as age or gender (Levitan, Mishra, and Bangalore 2016) information.The authors showed that audio, as compared to lexical, is used more successfully to predict gender. Hence, we hypothesize that a combination of these modalities leads to an increase in the leakage of the sensitive variable.

We train the six setups separately for each dataset described in Section 4.1 for activation and valence. We report the average across five-fold speaker-independent cross-validation in Table 1a and Table 1b. We find that:

- A network trained to only recognize emotion is generally discriminative for gender as well. For instance we obtain a leakage of 0.73 when training a multi-modal network for activation and of 0.64 when trained for valence on MuSE.
- In unimodal systems, leakage is higher when systems are trained using only audio streams compared to lexical.
- Leakage of gender in learned representation is higher for multimodal systems than that for the unimodal systems for both, MuSE and IEMOCAP (the two datasets with both audio and lexical information).

Our results suggest that models that aren't explicitly trained for gender recognition, or, that don't use gender as an input feature, still learn representations that are discriminative to identify gender. This leakage is more prominent when the input stream is audio as compared to lexical, but the leakage compounds in multimodal systems.

5.2 Question 2

Q: *How does the privacy metric change when a network is trained to preserve privacy?*

HYPOTHESIS: *Representations that are gender-invariant are less prone to leakage when attacked by an adversary, leading to better privacy preservation.*

Previous research has shown that obtaining a representation from a model trained to be invariant to gender, age, or location leads to better protection from an attacker who tries to recover this information (Coavoux, Narayan, and Cohen 2018). Previous research (Elazar and Goldberg 2018) has also shown that while the representations might be trained such that leakage of sensitive variable is reduced to chance, the attacker might still be able to recover this information. Hence, we concentrate on using this incapability as our primary metric. To test our hypothesis, we train the adversarial variants of the six models as mentioned above, while making sure that the leakage in the models is reduced to chance performance and compare our results to those in Table 1a and Table 1b. We train the multimodal models only for MuSE and IEMOCAP. Our results in Table 1a and Table 1b show that:

- The privacy metric is always higher when the representations are trained adversarially, compared to generally.
- Even when leakage is adversarially reduced to chance, the attacker is still able to recover information about gender.
- The privacy metric is in general always lower for audio than for lexical based unimodal systems.
- Multimodal systems often have the lowest privacy metric.

Our results suggest that, though the privacy of the learned representation is improved by reducing leakage while training, the attacker can still recover that information. This effect is especially compounded for multimodal systems. While previous work has concentrated on text (Section 2), our work shows how audio is the major culprit and that models involving audio as input are easier to exploit, even when trained adversarially for privacy preservation.

5.3 Question 3

Q: *How does emotion recognition performance change when networks are trained to preserve privacy?*

HYPOTHESIS: *There is a minor drop in emotion recognition performance when models are trained to preserve privacy.*

Previous research has shown that training a model invariant to a dataset variable might lead to drop in performance on

Table 1: Results using general (left) and privacy preserving models (right) for activation and valence prediction. U-UAR, U(M/F)-UAR for male/female, L-leakage, P-privacy metric, MI-membership identification. ***Bold-Italic*** shows significant improvement in metrics as compared to general classification model and *Italic* shows significant difference in metrics as compared to the privacy preserving model. Significance is established using paired t-test at adjusted p-value< 0.05.

(a) Prediction of activation using general (left) and privacy preserving classification (right)

		General Classification						Privacy Preserving Classification				
		L	U(M)	U(F)	U	P	MI	U(M)	U(F)	U	P	MI
Audio	Imp	0.69	0.65	*0.62*	*0.63*	0.35	0.71	0.64	0.57	0.60	***0.44***	0.68
	Pod	0.71	0.69	0.70	0.70	0.32	0.73	0.68	0.69	0.69	***0.44***	***0.68***
	Iem	0.73	0.66	0.69	0.67	0.30	0.72	***0.68***	0.70	***0.69***	***0.43***	***0.67***
	MuS	0.72	*0.61*	*0.64*	*0.63*	0.33	0.75	0.58	0.61	0.60	***0.45***	***0.69***
Lexical	Iem	0.62	0.51	0.52	0.52	0.39	0.59	***0.55***	***0.56***	***0.56***	***0.48***	***0.55***
	MuS	0.64	0.54	0.56	0.55	0.38	0.60	***0.58***	0.57	***0.58***	***0.47***	0.58
Multimodal	Iem	0.74	0.66	0.70	0.68	0.30	0.74	0.66	0.69	0.68	***0.41***	***0.67***
	MuS	0.73	0.65	0.66	0.66	0.31	0.76	0.65	0.64	0.65	***0.43***	***0.69***

(b) Prediction of valence using general (left) and privacy preserving classification (right)

		General Classification						Privacy Preserving Classification				
		L	U(M)	U(F)	U	P	MI	U(M)	U(F)	U	P	MI
Audio	Imp	0.56	0.53	0.49	*0.51*	0.44	0.70	0.51	0.49	0.48	***0.48***	0.68
	Pod	0.60	0.56	0.57	0.56	0.42	0.71	0.55	0.56	0.56	***0.47***	0.70
	Iem	0.62	0.60	0.61	0.60	0.39	0.70	0.60	0.62	0.61	***0.45***	0.68
	MuS	0.58	0.50	0.47	0.48	0.42	0.72	0.48	0.47	***0.47***	0.46	0.71
Lexical	Iem	0.61	0.64	0.65	0.65	0.41	0.62	***0.67***	***0.68***	***0.67***	***0.46***	0.62
	MuS	0.57	0.68	0.69	0.68	0.45	0.63	0.70	0.71	0.70	0.47	0.62
Multimodal	Iem	0.68	0.67	0.71	0.69	0.32	0.70	0.68	0.70	0.69	***0.45***	***0.68***
	MuS	0.64	*0.67*	0.66	*0.67*	0.38	0.71	0.64	0.65	0.65	***0.46***	0.71

the primary task, especially when there exists known correlations or biases in the datasets between the target label for the primary task and the secondary task (Meng et al.)

We compare the performance for predicting activation and valence of the models trained just to predict emotion (Act: Table 1a, Val: Table 1b) versus the model trained to enhance privacy while still predicting emotion in (Act: Table 1a, Val: Table 1b). Our results suggest that, in general there is no significant effect on the performance on the primary task when we train privacy preserving networks. We find that the performance is either maintained, e.g., Act: multimodal-MuSE; Val: multimodal-IEMOCAP, or there is a slight decrease in performance for some setups, e.g., Act: unimodal-acoustic-MuSE; Val: multimodal-MuSE. In multiple cases, such as Act/Val:unimodal-lexical-MuSE/IEMOCAP, contrary to some previous work, we also see a significant increase in performance, implying that making the model invariant to gender increases its robustness by not learning replicable associations between gender and emotion label.

5.4 Question 4

Q: *How does the adversarial component's strength impact emotion recognition performance and the privacy metric?*
HYPOTHESIS: *As the strength of the adversarial component increase, the privacy metric increases and the performance on the pimary task is unchanged.*
Our results in Section 5.2 suggest that while the leakage of the model was reduced to chance performance, the attacker is still capable of recovering this information. We analyze the effect of the strength of the adversarial component on the performance of the primary task and the privacy metric.

We find that the emotion recognition performance is generally unaffected with a change in λ, as expected from the results in Section 5.3. We observe that the the attacker is usually less capable of inferring gender from learned representations when $\lambda = 0.50$ as compared to when $\lambda = 0.75$. For example, the privacy metric for the unimodal audio system trained on MuSE increases from 0.39 to 0.45. But contrary to our expectation, we often see a decrease in the privacy metric when we move from $\lambda = 0.75$ to $\lambda = 1.00$ for both activation and valence. For example, the privacy metric for the unimodal audio system trained on MuSE decreases from 0.45 to 0.41. The decrease in the privacy metric as $\lambda \to 1$ could be attributed to overfitting of data (Schmidt et al. 2018) when being trained for invariance to the sensitive variable which the attacker network is able exploit. This suggests that an increase in the strength of the adversarial component doesn't necessarily correlate to an increase in the privacy metric.

5.5 Question 5

Q: *Focusing on gender, how does the performance of emotion recognition change when a network is trained to preserve privacy?*
HYPOTHESIS: *Learning representations invariant to gender will affect performance on the primary task in an imbalanced manner across subgroups.*
Previous research (Bagdasaryan and Shmatikov 2019) has

shown that training models invariant to race or gender can harm performance for one group more than others. This may be worrying when the prediction is used for sensitive application such as intervention or policing. Hence, we analyze if the performance on emotion recognition is affected in an imbalanced way for the models trained to enhance privacy.

We compare the performance for predicting activation and valence of the models trained just to predict emotion (Act: Table 1a, Val: Table 1b) versus the model trained to enhance privacy while still predicting emotion in (Act: Table 1a, Val: Table 1b). We find that while the performance is affected differently for the subgroups, the effect is not consistent across multiple setups and datasets. For example, the unimodal-acoustic system trained on MSP-Improv for activation classification decreases in performance for both the male and female groups, but the effect on the female group is greater. But the pattern isn't consistent across other datasets for the same model setup. Our takeaway from this analysis is cautionary, that though the privacy metric increases when a model is adversarially trained to enhance privacy, we need to ensure that the performance of the model on that dataset doesn't harm one subgroup more than the other.

5.6 Question 6

Q: *Does the location of the adversarial component within a network affect the privacy metric and emotion recognition performance?*
HYPOTHESIS: *Unlearning the demographic variable in separate pooled streams will improve the privacy metric.*
Previous work has shown that curtailing a variable on intermediate layers often leads to a difference in the performance of the classifier (Chabanne et al. 2017). As seen in Section 5.1, audio is more prone to leakage than lexical information, hence, a multimodal system's privacy metric might benefit from curtailing audio separately. Our initial multimodal model (Fig 1) only allows for the same strength and parameters of the adversarial component to be applied for both audio and lexical streams. To test our hypothesis, we place the same adversarial component after the mean pooling layer of both input streams, allowing us separate control of gender invariance for both modalities, before concatenation of representation.

We show our results in Table 2. We find that, using adversarial component separately for each input stream improves privacy metric for emotion recognition models trained on both datasets, as compared to using one adversarial component. This suggests that a granular control of invariance over modalities leads to better defense of representations against gender identification.

5.7 Question 7

Q: *Does the privacy preserving paradigm help defend against other attacks such as membership identification?*
HYPOTHESIS: *Membership identification will decrease when models are trained to be invariant to speaker.*
Membership identification is defined as an attack that tries to identify if samples from a speaker 'x' were present in the training set (Li et al.). (Papernot 2018) showed that removing identifying factors from learned representations reduces the probability of membership leakage. For this analysis, we ask two questions: (a) can we defend against membership identification using a proxy task and, (b) can we defend against both, gender and membership identification?

Table 2: Results for activation (Act) and valence (Val) prediction using multimodal input, when adversarially unlearning gender in each input (Priv-E) [left] stream separately. U-UAR, U(M/F)-UAR for male/female, P-privacy metric, MI-membership identification. ***Bold-Italic*** shows significant improvement in the privacy metric as compared to model trained to preserve privacy by maximizing loss on the concatenated representation (Priv-C) [right]. Significance is established using paired t-test, adjusted p-value< 0.05.

		Priv-E			Priv-C		
		U	P	MI	U	P	MI
Act	Iem	0.66	***0.43***	0.73	0.68	0.41	0.67
	MuS	0.65	0.44	0.74	0.65	0.43	0.69
Val	Iem	0.67	***0.46***	0.70	0.69	0.45	0.68
	MuS	0.66	0.47	0.74	0.65	0.46	0.71

We train an attack model for membership identification as specified in Section 4.2. We find that while adversarial removal of gender in the learned representation (Act: Table 1a and Val: Table 1b) does lead to reduced membership identification, as compared to a model trained solely for emotion recognition (Act: Table 1a and Val: Table 1b), the membership identification is still far higher than chance.

Our goal is to be unable to identify whether samples from speaker 'x' exist in the training set. This is different from the usual membership defense that prevents prediction of presence of a data-point pair $(input_x, output_x)$ is in the training set. As a result, we require a proxy task, because our model cannot use samples from the speakers not in the training set even to induce invariance. We hypothesize that given randomly chosen speakers from the population, speaker-invariant training leads to representations that are less likely to encode speaker-specific information. This will make it harder for the attacker to identify membership of a particular speaker in the training set. We train the emotion recognition models specified in Section 4.1 and replace the gender invariance sub-network with speaker invariance and use the same membership attack network.

We show our results in Table 3. We find that models trained to be invariant to speaker identity have significantly lower UAR for membership identification than those trained solely to recognize emotion, or trained invariant to gender, which matches our hypothesis.

Extension towards multi-attribute invariance. We train our emotion recognition model using both the adversarial components (speaker id and gender) and the primary classification task i.e., emotion recognition. This ensures that the model can defend against both, gender and membership identification attacks. We report our results in Table 3. We find that we can successfully train models that are safer against both, gender and membership identification attacks, while still maintaining similar performance on the primary

Table 3: Results for activation and valence prediction, for general classification (General), and, when adversarially unlearning subject identity (Priv-SubjectID) and both subject identity and gender (priv-Multiple). U-UAR, U(M/F)-UAR for male/female, P-privacy metric, MI-membership identification. ***Bold-Italic*** shows significant improvement in metrics as compared to general classification model and *Italic* shows significant difference in metrics as compared to the privacy preserving models. Significance is established using paired t-test at adjusted p-value< 0.05.

		General			Priv-SpeakerID			Privacy-Multiple		
		U	P	MI	U	P	MI	U	P	MI
					Activation					
Audio	Imp	*0.63*	0.35	0.71	0.59	***0.40***	***0.58***	0.59	***0.45***	***0.58***
	Pod	0.70	0.32	0.73	0.67	***0.37***	***0.60***	0.69	***0.46***	***0.59***
	Iem	0.67	0.30	0.72	0.66	***0.35***	***0.58***	0.67	***0.43***	***0.57***
	MuS	*0.63*	0.33	0.75	0.61	***0.36***	***0.62***	0.59	***0.44***	***0.60***
Lexical	Iem	0.52	0.39	0.59	0.51	0.40	0.52	0.53	***0.48***	***0.52***
	MuS	0.55	0.38	0.60	0.52	0.39	0.53	0.54	***0.47***	***0.52***
Multimodal	Iem	0.68	0.30	0.74	0.67	0.33	***0.58***	0.66	***0.40***	***0.57***
	MuS	0.66	0.31	0.76	0.65	0.33	***0.60***	0.65	***0.40***	***0.58***
					Valence					
Audio	Imp	*0.51*	0.44	0.70	0.47	0.45	***0.54***	0.47	0.48	***0.53***
	Pod	0.56	0.42	0.71	0.55	0.43	***0.56***	0.54	***0.48***	***0.56***
	Iem	0.60	0.39	0.70	0.61	0.41	***0.59***	0.60	***0.47***	***0.57***
	MuS	*0.48*	0.42	0.72	0.45	0.42	***0.60***	0.46	***0.46***	***0.58***
Lexical	Iem	0.65	0.41	0.62	0.67	0.41	***0.52***	0.66	***0.47***	***0.53***
	MuS	0.68	0.45	0.63	0.68	0.44	***0.53***	0.68	0.46	***0.53***
Multimodal	Iem	0.69	0.32	0.70	0.69	0.34	***0.57***	0.65	***0.43***	***0.56***
	MuS	*0.67*	0.38	0.71	0.64	0.37	***0.58***	0.62	***0.44***	***0.58***

task, as an evidence towards multi-attribute invariance.

6 Conclusion

In this work, we show how privacy preserving networks trained for emotion recognition can be used to protect against gender and membership identification. This provides a compelling case for separating the process of data processing on user devices and of task-specific training on central servers, thus increasing the privacy of the user. While in this paper we concentrate on a single primary task i.e., emotion recognition, this method can be extended to maximize utility on multiple primary tasks that are loosely related to each other and are benefited from a multi-task setup as shown for dialogue act and turn detection, and sentiment and topic classification (Ruder 2017).

For future work, we aim to explore how privacy enhanced representations can be learned for multiple primary tasks such as speaker verification and emotion recognition that may not be related to each other. This would enable us to deploy a generalized privacy model in form of SaaS which all developers could use the to obtain privacy enhanced representations that are then stored on the central server.

7 Acknowledgements

This material is based in part upon work supported by the Toyota Research Institute ("TRI"), the IBM PhD Fellowship Award, and by the National Science Foundation (NSF CAREER 1651740). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF, IBM, TRI, or any other Toyota entity.

References

Abadi, M.; Chu, A.; Goodfellow, I.; McMahan, H. B.; Mironov, I.; Talwar, K.; and Zhang, L. Deep learning with differential privacy. In *Proc. 2016 ACM Conference CCS*.

Aldeneh, Z.; Khorram, S.; Dimitriadis, D.; and Provost, E. M. 2017. Pooling acoustic and lexical features for the prediction of valence. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*.

Aron, A.; Melinat, E.; Aron, E. N.; Vallone, R. D.; and Bator, R. J. 1997. The experimental generation of interpersonal closeness: A procedure and some preliminary findings. *Personality and Social Psychology Bulletin* 23(4):363–377.

Bagdasaryan, E., and Shmatikov, V. 2019. Differential privacy has disparate impact on model accuracy. *arXiv preprint arXiv:1905.12101*.

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*.

Bolukbasi, T.; Chang, K.-W.; Zou, J. Y.; Saligrama, V.; and Kalai, A. T. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in neural information processing systems*.

Busso, C.; Bulut, M.; Lee, C.-C.; Kazemzadeh, A.; Mower, E.; Kim, S.; Chang, J. N.; Lee, S.; and Narayanan, S. S.

2008. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation* 42(4):335.

Busso, C.; Parthasarathy, S.; Burmania, A.; AbdelWahab, M.; Sadoughi, N.; and Provost, E. M. 2017. Msp-improv: An acted corpus of dyadic interactions to study emotion perception. *IEEE Transactions on Affective Computing*.

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}*.

Chabanne, H.; de Wargny, A.; Milgram, J.; Morel, C.; and Prouff, E. 2017. Privacy-preserving classification on deep neural network. *IACR Cryptology ePrint Archive*.

Chaplin, T. M. 2015. Gender and emotion expression: A developmental contextual perspective. *Emotion Review*.

Chollet, F. 2015. keras. https://github.com/fchollet/keras.

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

Corbett-Davies, S., and Goel, S. 2018. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023*.

Davidson, T.; Bhattacharya, D.; and Weber, I. 2019. Racial bias in hate speech and abusive language detection datasets. *arXiv preprint arXiv:1905.12516*.

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

Evfimievski, A. 2002. Randomization in privacy preserving data mining. *ACM Sigkdd Explorations Newsletter*.

Ganin, Y., and Lempitsky, V. 2014. Unsupervised domain adaptation by backpropagation. *arXiv:1409.7495*.

Gideon, J.; Schatten, H. T.; McInnis, M. G.; and Provost, E. M. 2019. Emotion recognition from natural phone conversations in individuals with and without recent suicidal ideation. In *The 20th Annual Conference of the International Speech Communication Association INTERSPEECH 2019*.

Gomez-Hidalgo, J. M.; Martin-Abreu, J. M.; Nieves, J.; Santos, I.; Brezo, F.; and Bringas, P. G. 2010. Data leak prevention through named entity recognition. In *IEEE Second International Conference on Social Computing*.

Hajian, S., and Domingo-Ferrer, J. A study on the impact of data anonymization on anti-discrimination. In *2012 IEEE 12th ICDM*.

Huang, C.; Zaiane, O.; Trabelsi, A.; and Dziri, N. 2018. Automatic dialogue generation with expressed emotions. In *Proceedings of NAACL*.

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.; 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*.

Kifer, D., and Machanavajjhala, A. 2011. No free lunch in data privacy. In *Proceedings of the ACM SIGMOD International Conference on Management of data*.

Levitan, S. I.; Mishra, T.; and Bangalore, S. 2016. Automatic identification of gender from speech. In *Proceeding of Speech Prosody*, 84–88.

Li, N.; Qardaji, W.; Su, D.; Wu, Y.; and Yang, W. Membership privacy: a unifying framework for privacy definitions. In *Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security*. ACM.

Lotfian, R., and Busso, C. 2017. Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings. *IEEE Transactions on Affective Computing*.

Matton, K.; McInnis, M. G.; and Provost, E. M. 2019. Into the wild: Transitioning from recognizing mood in clinical interactions to personal conversations for individuals with bipolar disorder. *Proc. Interspeech 2019* 1438–1442.

Meng, Z.; Li, J.; Chen, Z.; Zhao, Y.; Mazalov, V.; Gang, Y.; and Juang, B.-H. Speaker-invariant training via adversarial learning. In *2018 ICASSP*. IEEE.

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*.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, 3111–3119.

Papernot, N. 2018. A marauder's map of security and privacy in machine learning. *arXiv preprint arXiv:1811.01134*.

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

Rosenberg, A. 2012. Classifying skewed data: Importance weighting to optimize average recall. In *Thirteenth Annual Conference of the International Speech Communication Association*.

Ruder, S. 2017. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*.

Schmidt, L.; Santurkar, S.; Tsipras, D.; Talwar, K.; and Madry, A. 2018. Adversarially robust generalization requires more data. In *Advances in NIPS*.

Soto, J. A., and Levenson, R. W. 2009. Emotion recognition across cultures: The influence of ethnicity on empathic accuracy and physiological linkage. *Emotion* 9(6):874.

Tieleman, T., and Hinton, G. 2012. Lecture 6.5—RmsProp: Divide the gradient by a running average of its recent magnitude. COURSERA: Neural Networks for Machine Learning.

Zhao, H.; Chi, J.; Tian, Y.; and Gordon, G. J. 2019. Adversarial task-specific privacy preservation under attribute attack. *arXiv preprint arXiv:1906.07902*.