Multimodal Autoencoder: A Deep Learning Approach to Filling In Missing Sensor Data and Enabling Better Mood Prediction

Natasha Jaques, Sara Taylor, Akane Sano, and Rosalind Picard
Media Lab, Massachusetts Institute of Technology
Cambridge, Massachusetts 02139
Email: {jaquesn, sataylor, akanes, picard}@media.mit.edu

Abstract—**To accomplish forecasting of mood in real-world situations, affective computing systems need to collect and learn from multimodal data collected over weeks or months of daily use. Such systems are likely to encounter frequent data loss, e.g. when a phone loses location access, or when a sensor is recharging. Lost data can handicap classifiers trained with all modalities present in the data. This paper describes a new technique for handling missing multimodal data using a specialized denoising autoencoder: the Multimodal Autoencoder (MMAE). Empirical results from over 200 participants and 5500 days of data demonstrate that the MMAE is able to predict the feature values from multiple missing modalities more accurately than reconstruction methods such as principal components analysis (PCA). We discuss several practical benefits of the MMAE's encoding and show that it can provide robust mood prediction even when up to three quarters of the data sources are lost.**

1. Introduction

Affective Computing studies frequently collect rich, multimodal data from a number of different sources in order to be able to model and recognize human affect. These data sources — whether they are physiological sensors, smartphone apps, eye trackers, cameras, or microphones — are often noisy or missing. Increasingly, such studies take place in natural environments over long periods of time, where the problem of missing data is exacerbated. For example, a system trying to learn how to forecast a depressed mood may need to run for many weeks or months, during which time participants are likely to not always wear their sensors, and sometimes miss filling out surveys. While research has shown that combining more data sources can lead to better predictions [1], [2], as each noisy source is added, the intersection of samples with clean data from every source becomes smaller and smaller. As the need for long-term multimodal data collection grows, especially for challenging topics such as forecasting mood, the problem of missing data sources becomes especially pronounced.

While there are a number of techniques for dealing with missing data, more often than not researchers may choose to simply discard samples that are missing one or more modalities. This can lead to a dramatic reduction in the number of samples available to train an affect recognition model, a significant problem for data-hungry machine learning models. Worse, if the data are not missing completely at random, this can bias the resulting model [3].

In this paper we propose a novel method for dealing with missing multimodal data based on the idea of denoising autoencoders [4]. A denoising autoencoder is an unsupervised learning method in which a deep neural network is trained to reconstruct an input that has been corrupted by noise. In most cases, noise is injected by randomly dropping out some of the input features, or adding small Gaussian noise throughout the input vector. In contrast, we focus on the case where a whole block of features may go missing at one time — specifically, all of those features that are computed using the data from a single modality.

We demonstrate that by using a new model, which we call a Multimodal Autoencoder (MMAE), it is possible to accurately reconstruct the data from a missing modality, something that cannot be done with other techniques such as PCA. Further, we show that the MMAE can be trained with additional neural network layers designed to perform classification, effectively leveraging information from both unlabeled and labeled data. We present empirical results comparing MMAE to several other methods for dealing with missing data, and demonstrate that the MMAE consistently gives the best performance as the number of missing modalities increases.

Results are shown for the task of predicting tomorrow's mood, health, and stress, using data collected from physiological sensors, a smartphone app, and surveys. The goal of this research is to build a real-world system that can not only help participants predict their future mood and make adjustments to improve it, but also help detect early warning signs of depression, anxiety, and mental illness. However, the data inevitably contain samples with missing modalities, which can easily occur when a participant's smartphone cannot log data, or when sensor hardware malfunctions.

Previous work on this dataset (e.g. [5], [6], [7]) dealt with this problem by simply discarding samples for which any modality was missing. Therefore, these models cannot make accurate mood predictions if any of the data sources go missing. This is highly problematic if the models are to be used for any sort of real-world mental-health treatment and prevention program, as data frequently go missing during long-term use "in the wild".

978-1-5386-0563-9/17/$31.00 ©2017 IEEE

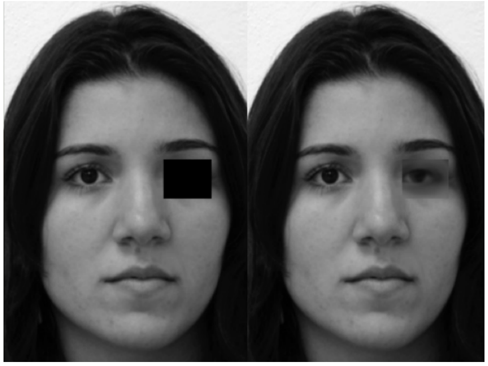

Figure 1. Image inpainting with an autoencoder, reproduced from [12]

In contrast, the new MMAE enables accurate mood prediction even with several missing modalities. Below we will show that in addition to being robust, the MMAE provides added benefits that may allow individuals with privacy or comfort concerns regarding the collection of certain types of data to opt out of providing such data, yet still enjoy the benefits of a mood forecasting system.

2. Related Work

Previous research has used autoencoders to enhance emotion recognition systems. Deng and colleagues demonstrate that an autoencoder can be used to improve emotion recognition in speech through transfer learning from related domains [8]. Xue and others use an autoencoder as a pretraining step in a semi-supervised learning framework to disentangle emotion from other features in speech [9]. A recent, related approach uses auto-encoders for both speech emotion classification and domain adaptation, taking advantage of their ability to learn from both labeled data and unlabeled data from other domains [10].

In the medical community, denoising autoencoders have been used to effectively compress data from large, sparse, extremely noisy Electronic Health Records (EHRs) into a much smaller embedding [11]. The authors show that the autoencoder embedding can drastically improve classification accuracy over the raw and noisy feature vector, or over other dimensionality reduction techniques such as PCA.

To the best of our knowledge, no previous work has proposed using autoencoders to fill in features from missing data sources. Some research that is conceptually similar to this idea comes from the computer vision community, which has investigated using autoencoders for the purpose of image inpainting [12], [13], [14]. In this problem, a large swath of an image has been removed or masked, and the task of the autoencoder is to hallucinate plausible values for the missing pixels based on related images it has seen in the training data (see Figure 1 for an example). This task is similar to our problem, because we consider the case when many related feature values go missing at once; for example, if the smartphone app encounters an error, we can no longer compute any of the many features relating to the participant's location, calls, or SMS. However, it should be noted that image inpainting may be a considerably easier task than filling in missing sensor data, because an inpainting autoencoder can take advantage of the strong spatial regularities of images and high correlations in values of neighbouring pixels that occur in natural images, not to mention the abundance of image data that exists for unsupervised learning.

3. Mood Prediction Dataset

The task at hand is to predict individuals' mood, health, and stress tomorrow night by using today's data about their physiology and behavior. The data we use were collected as part of a large-scale study of undergraduate students entitled SNAPSHOT: Sleep, Networks, Affect, Performance, Stress, and Health using Objective Techniques [15]. Rich, noisy, multimodal data was collected from 206 participants over 30 days each using wearable sensors, a smartphone app, and surveys. These data, along with weather information collected using DarkSky's Forecast.io API [16], were used to compute a total of 343 features. Only a brief overview of the data is provided here; for more details see [5], [6], [17].

Wrist-worn Affectiva Q sensors were used to collect 24-hour-a-day skin conductance (SC), skin temperature, and 3-axis accelerometer data, from which features such as step count, stillness, and SC responses (which relate to emotional arousal and stress) were extracted. Daily survey features included self-reported behaviors such as academic activities, exercise, and sleep. We include additional variables for day of the week, and whether it is a night before a school day. The smartphone app logged participants' calls, text messages, screen on/off events, and location throughout the day. In addition to extracting features about participants' communication and phone usage, location patterns were modeled with a Gaussian Mixture Model.

Each morning and evening, participants self-reported their mood (sad/happy), stress (low/high), and health (sick/healthy) on a scale from 0-100. Binary classification labels were assigned to the top and bottom 40% of these scores, discarding the middle 20% due to their questionable nature as either a 'happy' or 'sad' state, for example[1]. To predict future mood and wellbeing, features from today are combined to predict mood, stress, and health labels tomorrow night. All 5,547 days for which any data are present are divided into non-overlapping training, validation, and testing sets using a 65/20/15% split. Data from a single person may appear in multiple sets, to allow for comparison with previous work.

As with many Affective Computing studies, the multimodal, real-world nature of the dataset leads to inevitable problems with missing data, as Table 1 makes clear. While 206 participants × 30 days should lead to a total of 6180 days worth of data, there are only 5547 samples for which at least 40% of the features can be computed. This number is reduced significantly when we consider only those samples

[1] Note: this is an improvement from previous work [5], [6] in which the middle 40% of scores were discarded.

TABLE 1. MISSING DATA IN THE SNAPSHOT STUDY

Data	Num. samples	Percent
All days in the study	6180	100%
More than 40% clean data	5547	89.7%
All modalities present	3819	61.8%
Labeled	2951	47.7%
Labeled and all modalities	2886	46.7%

for which all of the multimodal data sources are available. The number of available samples drops even more precipitously when we must consider only those samples that have a supervised training label, especially when discarding the middle 20% of ratings[2].

If we wish to train a supervised learning model using only samples with all modalities, *we can use only half of the available data*. Meanwhile, valuable information contained in the remainder of the data goes to waste.

4. Method

An autoencoder is an unsupervised learning technique in which a deep neural network is trained to reproduce an input X based on the reconstruction error between X and the network's output X'; e.g. if using squared reconstruction error, the model would be trained to optimize the following loss function:

$$L(X, X') = \|X - X'\|^2 \tag{1}$$

A key feature of autoencoders is learning a useful representation of the data, often in a compressed format. The input $X \in \mathbb{R}^D$ must first be transformed into an *embedding* $Z \in \mathbb{R}^K$, often such that $K << D$; see Figure 2 for a graphical representation. The mapping from X to Z is accomplished by the *encoder* portion of the network. For example, if the encoder contains only a single neural network layer, then:

$$Z = \alpha(W_e X + b_e) \tag{2}$$

where W_e, b_e are the linear weights and bias and α is typically a non-linear activation function, for example a Rectified Linear Unit (ReLU).

The second half of the network, the *decoder*, maps Z to the reconstruction X'; i.e.:

$$X' = \alpha(W_d Z + b_d) \tag{3}$$

As a regularization technique, it is sometimes effective to tie the weights of the encoder and decoder, such that $W_d = W_e^T$.

The encoder can be considered a more complex, non-linear dimensionality reduction technique. In the simple case of a 1-layer encoder with no activation function and mean squared error (MSE) loss, the network behaves like PCA, learning to project the input in the span of the first K principle components of the data [18]. However, with multiple hidden layers and non-linear activation functions at each layer, the embedding can learn to encode complex, higher-level features. Thus, the embedding Z can capture important conceptual information about the input data [19].

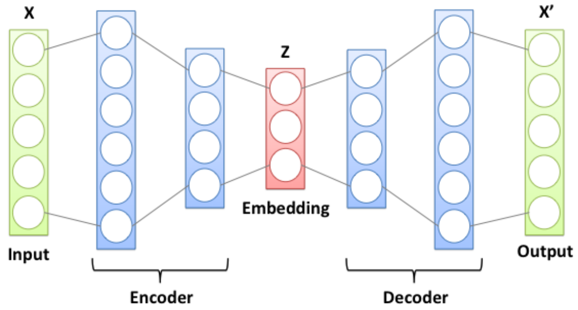

Figure 2. Autoencoder model

In a denoising autoencoder (DAE), the input X is corrupted with noise to obtain $\widetilde{X}$. The DAE is trained to reconstruct the original, noise-free input X from $\widetilde{X}$. Typically, the added noise takes the form of: a) Gaussian noise, $\widetilde{X}|X \sim \mathcal{N}(X, \sigma^2 I)$; b) masking noise, where a random fraction of the elements of X are set to 0; or c) salt and pepper noise, where a random fraction of the elements of X are set to their minimum or maximum value [4].

4.1. MMAE

The MMAE was developed to ameliorate the likely problem where a number of contiguous features from the same modality go missing at once. We start by normalizing all of the features to be in the range $[0, 1]$. We then represent a missing modality by filling all features from that modality with the special value -1. It is important to use a special value to indicate missing data that must be filled, rather than fill with a value such as 0 which could actually occur in the real data. To train the MMAE, we first use samples that have data from every modality to provide the ground truth noise-free X. At training time, for every sample X, we compute $\widetilde{X}$ by adding noise using two methods. First, we add simple masking noise to 5% of the features, as in [4]. Second, we randomly select one or more modalities and set all of the feature values for that modality to -1; essentially, masking entire modalities at once. The model is then trained to reproduce X from $\widetilde{X}$. Effectively, this means that the model must learn to predict reasonable values for the missing modality from the rest of the features. For example, it may use the participant's physiology and location patterns to predict her survey responses, such as how much time she spent in class, or whether she drank caffeine.

After training the autoencoder portion of the network with the clean, unsupervised examples for which all sensors are available, we then begin a second phase of training for classification. Here we connect the encoder to additional

[2] The number of samples with all modalities present overlaps more heavily with labeled samples in this dataset than is typical of most datasets, since the labels are collected from a survey, and other information from this survey is considered to be one of the modalities.

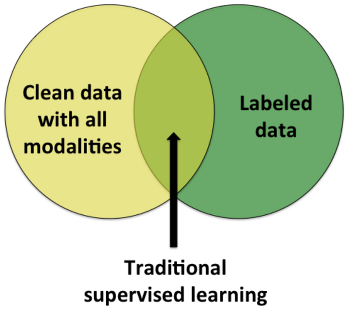
Clean data with all modalities
Labeled data
Traditional supervised learning

Figure 3. Data loss in traditional supervised learning paradigm

classification layers used for predicting mood, health, and stress. We allow gradients to backpropagate through the entire network, from the classification layers into the encoder. In this second phase, although we continue to add noise in the same way, we use all of the training data for which a label is available, whether it has data for every modality or not. As is presented in Figure 3, traditional supervised learning is only able to learn from the intersection of samples which are both clean and labeled. In contrast, the weights of the MMAE's encoder learn from both clean, unsupervised data with no labels, and noisy, supervised data with missing modalities, leveraging as much of the available data as possible.

We identify 11 modalities within the data, as in [7]; these are shown in Table 2. Note that physiology is sub-divided into features from four different time intervals during the day in order to ensure each modality has a roughly similar number of features. This could allow the MMAE to more easily predict an individual's physiology in the afternoon from her physiology in the morning. However, we believe this to be a realistic scenario, since often a participant will choose not to wear the sensor for only part of a day, e.g., if s/he has to participate in an extra-curricular activity such as a dance recital or swim meet.

Still, it is possible for multiple modalities to go missing at once, e.g. all four physiology modalities. Previous research has shown that denoising autoencoders are most effective when the noise injected during training matches the actual noise in the data distribution [14]. Therefore, we assessed the training data to determine how frequently each modality goes missing, and which modalities frequently go missing together. We found that in the SNAPSHOT data, the location modality is lost most frequently (likely due to participants disabling location services on their phone), and the second most likely pattern is that all of the smartphone app modalities (location, call, SMS, and screen) go missing together. We used this learned distribution over missing modalities to improve the training of the MMAE; we call this approach training with *structured noise*.

4.2. Implementation and Experiments

While using MSE is easy and most common, we found that using a cross-entropy (CE) reconstruction loss reliably led to better results for the MMAE than using MSE. The CE loss to be minimized is:

$$L_H(X, X') = -\sum_{k=1}^{D}[X_k \log X'_k + (1 - X_k)\log(1 - X'_k)]$$

Since cross-entropy is appropriate for binary values, before applying this loss we first normalized all of our features to the [0,1] range.

In addition, we experimented with implementing the MMAE as a Variational Autoencoder (VAE) [20], which constrains the features in the embedding to follow K independent Gaussian distributions. This makes it more likely that a random embedding sampled from a K-dimensional multivariate Gaussian with mean 0 and variance 1, will actually correspond to a plausible sample when passed through the decoder; in other words, it makes it possible to generate new samples by interpolating in the embedding space. While this ability to generate realistic-looking samples of data is interesting, we conducted experiments using the VAE version of our MMAE and found it did not improve reconstruction or classification performance.

To assess the MMAE, we compared it to two other dimensionality reduction techniques: PCA, and a supervised feature selection technique in with the features with the highest ANOVA F-value with the classification label in the training data were selected. We constrained each method to reduce the original 343 features to 100 dimensions to enable fair comparison; this allowed the PCA to capture 98% of the variance in the data, assuring a fair comparison. We also compared MMAE to four ways of dealing with missing data, including discarding the data when training the model, filling it with a special value like -1, filling it with the average for that feature, and filling it using a PCA reconstruction. PCA reconstruction of missing data was conducted by applying the inverse transformation learned by PCA to the 100-dimensional principle components vector.

We also compared the MMAE's classification performance to three other machine learning algorithms including Support Vector Machines (SVM), Logistic Regression (LR), and a feedforward neural network (NN). For all models we performed a grid search over possible hyperparameter settings, optimizing for performance on the validation set.

Due to space constraints we cannot report the optimal hyperparameters for every combination of model and classification label, but we can indicate that the MMAE autoencoder architecture that produced the lowest reconstruction error was: hidden layers of size [300,100] for the encoder, identical structure with tied weights for the decoder, softsign activation function, no dropout, and an L2 weight regularization coefficient of .001. The MMAE architecture that produced the best classification accuracy had hidden layers of [300,100] with tied weights for the autoencoder, classification layers of [50,20], ReLU activation and dropout throughout, and an L2 weight regularization coefficient of .01 for the autoencoder, but 0 for the classification layers.

All of the Tensorflow code developed to implement the MMAE – as well as the supporting algorithms, feature

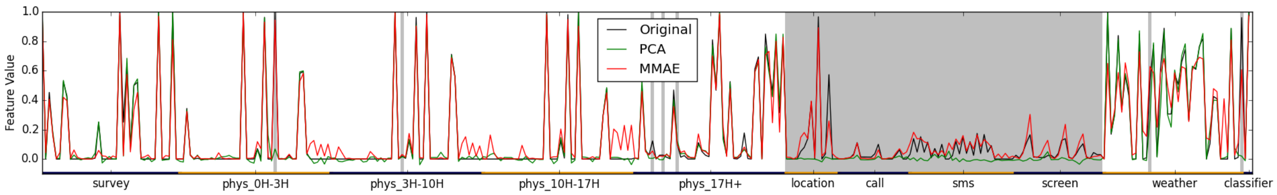

Figure 4. The full feature vector containing 11 modalities. MMAE reconstruction (red) and PCA reconstruction (green) are compared to the original data (black). Areas shaded grey have been masked to produce $\widetilde{X}$.

selection methods, and the hyperparameter search – has been open sourced and is available at https://github.com/natashamjaques/MultimodalAutoencoder.

5. Results

We first assess the ability of the MMAE to fill in missing modality data. As a comparison, we also reconstruct missing data with a PCA mapping learned on the training data. The PCA was able to explain 97.81% of the variance in the data when projecting down to 100 dimensions, indicating that it provides a strong baseline. In fact, PCA produces a more faithful reconstruction of the clean data than the MMAE; PCA obtains a Root Mean Squared Error (RMSE) of 0.036 on reconstructing clean test set data with no missing modalities, while MMAE scores 0.084.

However, the strength of the MMAE is its ability to restore missing modalities by predicting appropriate values based on the rest of the feature vector and similar patterns in the training data. Figure 4 shows that even in the case when four of the modalities go missing at the same time, the MMAE trained with structured noise is still able to predict realistic values for the features that have been masked with -1. In contrast, the PCA reconstruction hovers around whatever value was used to fill the missing data; in this case we chose a value of 0 to make it a fair comparison (since 0 is a frequent value in the real data, the RMSE will be lower if the PCA reconstruction does not differ much from the fill value), but still find that PCA is unable to reconstruct the missing features.

The difference in reconstruction performance between the MMAE and PCA is even more evident in Figure 5, which shows a close-up of reconstructed data from a single missing modality. PCA is able to recover one or two of the features, likely because they are highly correlated with other features in the vector which are not masked, and are thus redundant. However, in general PCA fails to reconstruct the missing data and again produces output hovering around 0. Conversely, the MMAE is able to accurately predict the missing feature values based on patterns learned in the training data, effectively restoring much of the original data.

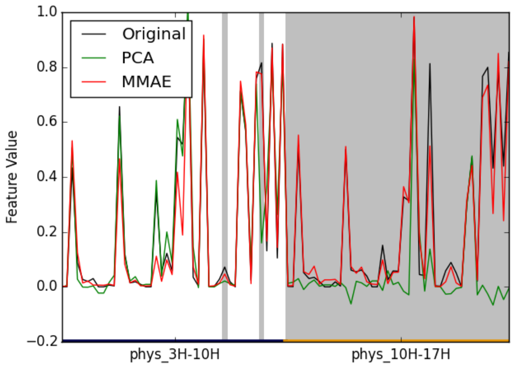

Figure 5. MMAE reconstruction (red), PCA reconstruction (green), original data (black). As in Figure 4, masked data has been shaded grey.

TABLE 2. RMSE FOR EACH MODALITY. BOLDED ENTRIES ARE SIGNIFICANT IMPROVEMENTS.

Modality	**Size**	**PCA**	**MMAE**	t
Survey	39	0.363	**0.263**	26.5
Physiology 12am-3am	43	0.319	**0.095**	84.9
Physiology 3am-10am	43	0.320	**0.086**	103.7
Physiology 10am-5pm	43	0.301	**0.091**	96.4
Physiology 5pm-12am	43	0.320	**0.093**	85.3
Location	15	0.590	**0.110**	133.8
Call	20	0.280	**0.044**	137.4
SMS	30	0.481	**0.078**	154.6
Screen	25	0.423	**0.081**	149.3
Weather	40	0.488	**0.253**	82.4
Day of week, school night	2	0.634	**0.276**	12.3
Total	343	0.411	**0.134**	104.5

5.1. Ability to reconstruct each modality

To test the MMAE's ability to reconstruct data from each of the different sources, each modality was dropped out over the entire test set, and this data was reconstructed with either an MMAE trained with by uniformly masking different modalities, or with PCA. It is clear from Table 2 that the MMAE produces decidedly lower RMSE when reconstructing data from a missing modality than PCA. A series of t-tests with Bonferroni correction were conducted to determine if MMAE produced significantly lower RMSE than PCA; all of the tests were significant at the $p = .001$ level.

From Table 2, it is interesting to note that the MMAE can more easily predict a person's physiology and behavioral patterns (e.g. call, sms, screen, etc.) than predict extrinsic factors like the weather or the day of the week. In particular,

TABLE 3. MOOD PREDICTION ACCURACY ON HELD-OUT TEST SET

Label	Model	Fill avg.	Fill -1	Feat. sel.	PCA	MMAE
Mood	LR	59.3	59.2	60.2	57.0	60.2
	SVM	61.8	59.5	61.2	59.3	59.1
	NN	60.3	58.2	60.9	62.5	61.5
Health	LR	59.7	59.8	57.9	56.7	58.9
	SVM	60.5	61.0	64.2	61.6	64.1
	NN	64.3	62.5	59.3	60.4	61.5
Stress	LR	62.5	61.7	59.3	59.2	60.3
	SVM	65.5	61.8	62.6	59.5	58.7
	NN	63.9	59.8	60.5	63.2	62.2

the RMSE for day of the week may be quite high because it is not possible to distinguish between similar week days; i.e. a student's physiology and location patterns may look the same whether it is Monday or Tuesday.

5.2. Using the MMAE embeddings for classification

We also tested the ability of the MMAE to produce embeddings that can be used effectively for classification. To do this each feature vector X was passed through the encoder to produce an embedding Z, then the embedding was used with other classifiers such as SVM. These results are compared to those obtained by applying other methods for dimensionality reduction or dealing with missing data; namely, PCA, feature selection, and filling the missing values with either the average or a special value like -1. Although some studies have reported dramatic improvement in prediction accuracy using autoencoder embeddings (e.g. [11]), in this case the MMAE embeddings did not approve classification performance above the comparison methods. Table 3 shows that the accuracy[3] in predicting mood, stress, and health on the held-out test set when using the embeddings is similar to that obtained with the other methods. A McNemar test [21] applied revealed no significant differences. The lack of improvement is likely due to the fact that the dataset is relatively clean (only about 30% of the supervised training examples contain noise). Further, the original feature vector used in the work of Miotto and colleagues contained 100,000s of extremely noisy features [11], whereas the 343 features from the SNAPSHOT data are already the result of several years worth of careful feature extraction, design, and selection based on domain knowledge, and are therefore already compressed and cleaned. Still, the embedding provides equivalent performance while compressing the data representation even further for enhanced computational efficiency. In addition, the embedding provided by the MMAE is a de-identified representation of otherwise highly sensitive and personal data, which may provide protection for privacy as long as the decoder is kept private.

5.3. Robust prediction with missing modalities

The most important use case of the MMAE is to be able to deal effectively with real-world noisy data in which several modalities may go missing at once. Therefore, we compared the MMAE to several other methods for dealing with missing data: simply discarding it and training only on clean samples, filling it with a special value like -1, or performing PCA. Each of these methods are used to train a NN, as it was shown to give the best performance for mood and stress forecasting. The MMAE can directly make predictions using the additional classification layers connected to the encoder, as described in Section 4.1.

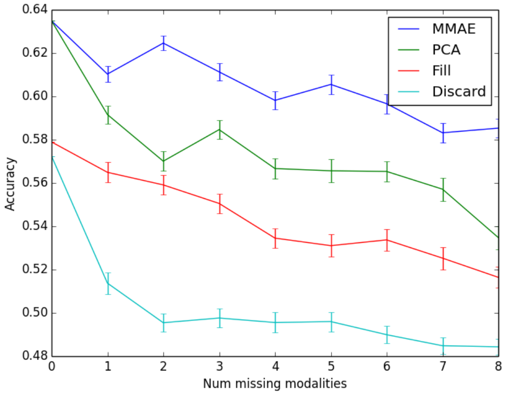

Figure 6. Stress prediction accuracy on the held-out test set as a function of the number of missing modalities. Error bars show 95% confidence intervals. Mood and health showed a similar pattern.

Figure 6 shows the performance of each of these methods on the test data as the number of modalities missing from the data increases. Note that the discard model was trained once on all available clean data, while the rest of the models were re-trained each time on training data with the appropriate number of missing modalities per row. As is obvious from Figure 6, the discard model — which represents the previous state-of-the-art for dealing with missing modalities in this dataset [5], [6], [7] — performs extremely poorly as more modalities go missing. This is likely to reflect the performance that can be expected from such a model when applied "in the wild" in a mood prediction app. Performance is slightly higher when the NN is trained on noisy data, but in general both dimensionality reduction methods (PCA and MMAE) give higher performance from the beginning, likely because they reduce the risk of overfitting. When the dataset is relatively clean, as is the case with the SNAPSHOT data, the MMAE may not provide a significant performance improvement over PCA. However, as the number of modalities lost increases, the MMAE reliably outperforms PCA, maintaining its mood prediction accuracy even when nearly three quarters of the original features are missing. Thus, we see that the new MMAE provides an important performance advantage for a real-world system in which multiple modes of data are sporadically present.

[3] We also compute Area Under the Curve (AUC) scores; they are extremely similar to the accuracy scores due to the balanced nature of the classification labels, and are not shown due to space constraints.

6. Discussion and Conclusion

We have described a new method for restoring missing sensor data, which is frequently lost in multimodal, real-world data collection settings. Empirical results demonstrate that the MMAE can accurately reproduce data from a lost modality, while other methods such as PCA cannot. The MMAE offers valuable new advantages for Affective Computing researchers who would like to train unbiased models on noisy data, accurately cluster noisy samples, or make robust predictions in the face of real-world data loss.

The MMAE has potential benefits in terms of providing enhanced flexibility and privacy to users of a mood prediction system. Because it can make accurate mood predictions even when data is lost, it could allow users to opt-out of providing data for all modalities. This could be particularly enticing to certain users, e.g. those who are uncomfortable wearing sensors throughout the day, or those who are concerned about privacy issues surrounding location data.

The MMAE also provides an effective feature reduction method that may enhance privacy; the embeddings learned by the MMAE can be used to provide roughly equal classification performance to the raw features, meaning that the raw features would not have to be stored once the embeddings are computed. The embeddings could potentially allow the highly sensitive personal data collected from this study to be shared with other researchers in a non-identifiable way.

We believe the MMAE provides an advance in the modeling of real-world mood prediction systems based on long-term multimodal data streams. Unlike prior methods, the MMAE is able to leverage valuable information from all available data, whether labeled, unlabeled, noisy, or clean. We have shown that the performance of machine learning models trained without considering missing data quickly deteriorates with data loss; however the MMAE's performance is relatively maintained even with significant loss of data. While models trained to account for missing data cannot provide reliable prediction performance as the level of noise increases, the MMAE can maintain its ability to predict tomorrow's mood even in the realistic situation where there is intermittent missing input data.

Acknowledgments

We would like to thank Dr. Charles Czeisler, Dr. Elizabeth Klerman, Conor O'Brien and other SNAPSHOT project members for their help in running the SNAPSHOT study. This work was supported by the MIT Media Lab Consortium, NIH Grant R01GM105018, Samsung Electronics, and Canada's NSERC program.

References

[1] Sidney K DMello and Arthur Graesser, "Multimodal semi-automated affect detection from conversational cues, gross body language, and facial features," *User Modeling and User-Adapted Interaction*, vol. 20, no. 2, pp. 147–187, 2010.

[2] Ashish Kapoor and Rosalind W Picard, "Multimodal affect recognition in learning environments," in *Proceedings of the 13th annual ACM international conference on Multimedia*. ACM, 2005, pp. 677–682.

[3] Andrew Gelman and Jennifer Hill, "Missing-data imputation," *Behavior research methods*, vol. 43, no. 2, pp. 310–30, 2007.

[4] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *Journal of Machine Learning Research*, vol. 11, no. Dec, pp. 3371–3408, 2010.

[5] Natasha Jaques, Sara Taylor, Asaph Azaria, Asma Ghandeharioun, Akane Sano, and Rosalind Picard, "Predicting students' happiness from physiology, phone, mobility, and behavioral data," in *Affective computing and intelligent interaction (ACII), 2015 international conference on*. IEEE, 2015, pp. 222–228.

[6] Natasha Jaques, Sara Taylor, Akane Sano, and Rosalind Picard, "Multi-task, multi-kernel learning for estimating individual wellbeing," in *Proc. NIPS Workshop on Multimodal Machine Learning, Montreal, Quebec*, 2015, vol. 898.

[7] Natasha Jaques, Sara Taylor, Ehimwenma Nosakhare, Akane Sano, and Rosalind Picard, "Multi-task learning for predicting health, stress, and happiness," in *Proc. NIPS Workshop on Machine Learning for Healthcare, Montreal, Quebec*, 2015.

[8] Jun Deng, Zixing Zhang, Erik Marchi, and Bjorn Schuller, "Sparse autoencoder-based feature transfer learning for speech emotion recognition," in *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*. IEEE, 2013, pp. 511–516.

[9] Wentao Xue, Zhengwei Huang, Xin Luo, and Qirong Mao, "Learning speech emotion features by joint disentangling-discrimination," in *Affective Computing and Intelligent Interaction (ACII), 2015 International Conference on*. IEEE, 2015, pp. 374–379.

[10] Jun Deng, Xinzhou Xu, Zixing Zhang, Sascha Frühholz, and Björn Schuller, "Universum autoencoder-based domain adaptation for speech emotion recognition," *IEEE Signal Processing Letters*, vol. 24, no. 4, pp. 500–504, 2017.

[11] Riccardo Miotto, Li Li, Brian A Kidd, and Joel T Dudley, "Deep patient: An unsupervised representation to predict the future of patients from the electronic health records," *Scientific reports*, vol. 6, 2016.

[12] O Shcherbakov and V Batishcheva, "Image inpainting based on stacked autoencoders," in *Journal of Physics: Conference Series*. IOP Publishing, 2014, vol. 536, p. 012020.

[13] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros, "Context encoders: Feature learning by inpainting," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2536–2544.

[14] Junyuan Xie, Linli Xu, and Enhong Chen, "Image denoising and inpainting with deep neural networks," in *Advances in Neural Information Processing Systems*, 2012, pp. 341–349.

[15] A. Sano, *Measuring College Students Sleep, Stress and Mental Health with Wearable Sensors and Mobile Phones*, Ph.D. thesis, MIT, 2015.

[16] The Dark Sky Company LLC, "Dark sky forecast api," 2016.

[17] A. Sano et al., "Prediction of happy-sad mood from daily behaviors and previous sleep history," in *EMBC*. IEEE, 2015.

[18] Yoshua Bengio et al., "Learning deep architectures for ai," *Foundations and trends® in Machine Learning*, vol. 2, no. 1, pp. 1–127, 2009.

[19] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[20] Diederik P Kingma and Max Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[21] Omolola A Adedokun and Wilella D Burgess, "Analysis of paired dichotomous data: A gentle introduction to the mcnemar test in spss," *Journal of MultiDisciplinary Evaluation*, vol. 8, no. 17, pp. 125–131, 2011.