Towards an Estimation of Internal State Through Dense, Multi-Modal Representation Learning

Matthew J. Vowels
University of Surrey
Centre for Computer Vision, Speech and Signal Processing
Guildford, Surrey, UK
m.j.vowels@surrey.ac.uk

I. Research Vision

The NHS spent £11.6 billion on mental health in 2016-17 [1] and 50% of all people will suffer from at least one mental disorder in their lifetime [2]. However, reports suggest that current mental treatments may not be particularly effective. For example, meta-analyses indicate that whilst the spontaneous recovery rate for depression is approximately 50% [3], the recovery rate for people who undergo therapy is also only 50% [1]. The field of mental health research (principally psychology) and the practice of mental health treatment therefore represent vital areas of research with vast room for improvement.

One burgeoning area of research involves the use of Artificial Intelligence (AI) and Machine Learning (ML) for applications in mental health and automated or online therapy [4]–[7]. Such applications include the estimation of mental health problems from various individual or combined modalities such as speech, vision, or text [8]. However, many of these applications (a) apply ML to specific tasks, such as depression, and (b) reply on supervised datasets which are expensive and time-consuming to collect.

Unsupervised methods represent a means to learn embeddings which are broadly useful across a range of downstream tasks without requiring expensive, high quality annotations. Such methods are highly attractive in view of the abundance of unlabelled data in the wild. Unfortunately, unsupervised methods currently under-perform their supervised counterparts on specific tasks, but narrowing this performance gap is a goal of the ML research community [9].

We make the assumption that the internal state and mental health of an individual can be modelled as a set of latent variables to be inferred from unlabelled, observational data (e.g. in the form of video and/or audio and/or text). The structure of these latent variables is not assumed *a priori* although literature debates the associations between mental health and various patterns in internal states such as emotion, valence and arousal [10], [11]. Such associations may be also be highly complex and dynamic, and therefore benefit from being modelled with state-of-the-art ML techniques.

It is common in machine learning to seek to minimize a pre-defined loss function that represents a metric for model fit, such as classification accuracy. If no causal constraints are imposed on the model, it will leverage all available correlation in order to minimize this loss function without regard for the true, underlying generative structure. However, as the old adage goes, correlation does not imply causation, and a high classification accuracy by itself may actually compromise reproducibility, robustness, and generalizability [12]. For rich datasets (datasets with many variables but relatively few datapoints), it may even be the case that *'correlation does not even imply correlation'* [13]. Furthermore, such naive models may be limited in their capacity to inform us of the relationships between underlying phenomena [12], [14], a task known as parameter estimation [15]. In the application of more traditional, parametric statistical models, we may benefit from model interpretability, but the models may suffer from misspecification, which may also lead to false positives, confounding, poor reproducibility, and biased parameter estimates [16]. We take an approach strongly responsive to theoretic discovery and domain knowledge to ensure that our model closely reflects the true underlying generative data structure whilst avoiding strong and restrictive assumptions that limit the expressive power of the model. In other words, we wish to 'let the data speak' [15].

Our research vision is therefore to develop unsupervised ML models capable of inferring the internal state of an individual over time, from multi-modal, observational data. The model should provide a densely informative and meaningful embedding over multiple time-scales, useful across a range of downstream tasks relevant to mental health. Such tasks include the prediction of transient states of emotion, valence, and arousal, as well as longer-term traits such as personality type and mental health disorders. The project hopes to contribute new knowledge to the fields of representation learning, psychology and mental health, as well as to provide useful tools for therapists and to help automate and standardise treatment. We acknowledge the current reproducibility and replication crises in the field of psychology [17], [18] and take an open and exploratory perspective to modelling the complex phenomena associated with the human psyche.

II. Recent Work

Our project involves researchers from both the Centre for Vision, Speech, and Signal Processing as well as the School of Psychology at the University of Surrey. We are undertaking two streams of research in parallel: One concerns the development of novel representation learning methods, and the other concerns psychological research relevant to validate and inform the development of these methods. We will now review the work to date.

A. Disentanglement

One facet of an approach such as ours is the goal of disentanglement, which is the idea of decomposing latent generative factors into their principle independent components. A review of the representation learning literature highlights the importance of achieving disentangled representations and is a long standing goal for the machine learning sciences. A disentangled representation allows for causal reasoning [19]–[21], fair machine learning [22]–[24], generalizability [25], [26], attribute transfer [27], [28], and improved performance on downstream tasks [29], [30]. These attributes are all relevant to our application focus of estimating internal state. Unfortunately, there are various identifiability issues associated with the unsupervised learning of disentangled representations which mean that some level of supervision is required.

Our initial work [31] (accepted to 15th IEEE Conference on Automatic Face and Gesture Recognition) sought to incorporate weak-supervision in the form of paired data, where the pairing corresponded to some shared latent attribute. The method was demonstrated to successfully disentangle subsets of latent factors according to this weak supervision, as well as learning an embedding which is more informative for downstream tasks (such as emotion prediction) than existing, unsupervised disentanglement methods [31], [32]. Such work is useful in providing a means to disentangle the facial attributes relevant to emotion prediction from static images.

Subsequent work [33] (accepted to the Conference on Computer Vision and Pattern Recognition, 2020) modified the initial work to learn representations that are invariant to confounding factors. Such work is important in an age where systemic bias associated with the use of ML algorithms is impacting individuals [34]–[36]. In developing methods for mental health applications, we need to be mindful that our algorithms are minimally affected by confounding or culturally and politically sensitive factors deriving from the data with which they are presented. When compared with existing methods, the work was shown to improve the prediction of biological sex from static images of faces in terms of classification score whilst being invariant to race.

B. Work on the Dynamics of Psychological Phenomena

The work described in the previous section explored both disentanglement and invariance for the learning of informative embeddings from static images of faces. However, we have alluded to the dynamic nature of psychological phenomena, and therefore intend to extend these methods to the time domain. Unfortunately, comparatively little is known about how psychological phenomena change from moment to moment, particularly in terms of high-level constructs such as emotion [37]. We developed open-source software enabling participants to rate their valence as they reviewed videos of both positive and negative interactions with a close partner. Such software will provide our collaborators in the psychology department with a tool to undertake research that provides insights into the dynamics of valence between partners. The software is described in an article accepted to the Journal of Counseling Psychology and available upon request [37].

Subsequent work sought to understand the extent to which self-report moment-to-moment ratings of valence are predictive of relationship outcomes. We took various forms of the cross-power-spectral-density estimate for the ratings from the two partners and trained a distribution of random forest classifiers using Leave-One-Out Cross-Validation to provide a distribution of accuracy scores for whether the couple broke up after two-years. Our initial hypothesis was that emotion ratings from two years prior would not be predictive of breakup, but our results indicate that certain features relating to the emotion dynamics are informative. Our results have been accepted for presentation at the International Association for Relationship Research (IARR) conference. Our results indicate that dynamics and interactions between individuals may provide useful signal, and this information will inform the development of our representation learning methods in future work.

III. Working Plan & Challenges

We have briefly described the two parallel streams of work to date, the first relating to the representation learning models themselves, and the second to the psychological research intended to inform, validate and apply these models. In the case of the former, there are two current restrictions: Firstly, the models assume static, independently and identically distributed data, and secondly, the models are concerned with the visual modality. Currently, we are developing an extension to the existing models for sequential data from the visual modality. This work is expected to be completed by June-July, 2020. Pursuant to the development of these sequential models, we will then extend them to handle the audio and textual modalities. We foresee the audio modality extension to be complete by the end of 2020, and the textual modality extension by mid 2021.

Once the models have been extended to handle sequential data across all three observational modalities, we can investigate how the representations from these modalities can be fused and integrated across a temporal hierarchy. We predict these temporal hierarchies to be useful on the basis that certain modalities reflect change at different temporal scales. For instance, the meaning of a sentence may be inferred across its spoken duration, which may be in the order of tens of seconds. In contrast, instantaneous facial expressions may change in the order of hundreds of milliseconds.

Learning representations for states that change across different temporal scales, as well as identifying a way to fuse representations derived from different modalities (each of which change across temporal scales in different ways) represents a significant engineering challenge. Furthermore, we wish to retain fairness and generalisability. ML algorithms are at the mercy of the quality of the data with which they are trained and the quality of the infused domain knowledge. Given the intended application of these algorithms to psychology and mental health, we must endeavour to design models which are invariant to culturally and politically sensitive factors. We foresee disentanglement and the sourcing/curation of balanced datasets to be the biggest aids in this regard.

In parallel with the model development described above, we will continue to collaborate with the School of Psychology. As yet, we have only applied one of our models to the prediction of the 6 prototypical emotions, and another model to the prediction of biological sex. Moving forwards, we wish to utilise our models to predict mental health related tasks, such as therapy outcomes and diagnoses. We will continue to investigate the role of interpersonal dynamics and interactions, given that recent work has highlighted their importance even for temporally distant outcomes (specifically two year break up).

The outcome of the project therefore falls in line with the goal presented in the first section: Un- or weakly-supervised model(s) which provide meaningful, fair, and densely informative embeddings derived from multiple modalities over multiple time-scales, which are predictive of a broad array of tasks relevant to psychology and mental health.

References

[1] NHS England, “Mental health five year forward view dashboard,” *Quarter 1 and 2*, 2018.

[2] Z. Steel, C. Marnane, C. Iranpour, T. Chey, J. W. Jackson, V. Patel, and D. Silove, “The global prevalence of common mental disorders: a systematic review and meta-analysis,” *International Journal of Epidemiology*, vol. 43, no. 2, pp. 476–493, 2014.

[3] H. Whiteford, “Estimating remission from untreated major depression: a systematic review and meta-analysis.,” *Psychological Medicine*, vol. 43, no. 8, pp. 1569–85, 2013.

[4] R. Cummins, M. P. Ewbank, A. Martin, V. Tablan, A. Catarino, and A. Blackwell, “TIM: a tool for gaining insights into psychotherapy,” *WWW*, 2019.

[5] M. P. Ewbank, R. Cummins, and V. Tablan, “Quantifying the association between psychotherapy content and clinical outcomes using deep learning,” *JAMA Psychiatry*, vol. 77, no. 1, pp. 35–43, 2020.

[6] S. Jaiswal, M. Valstar, K. Kusumam, and Greenhalgh, “Virtual human questionnaire for analysis of depression anxiety and personality,” *Proc. 19th ACM Intl. Conf. Intelligent Virtual Agents*, 2019.

[7] J. F. Cohn, N. Cummins, J. Epps, R. Goecke, J. Joshi, and S. Scherer, “Multimodal assessment of depression from behavioral signals,” *Handbook of Multimodal-Multisensor Interfaces*, vol. 2, pp. 375–417, 2018.

[8] A. B. R. Shatte, D. M. Hutchinson, and S. J. Teaague, “Machine learning in mental health: a scoping review of methods and applications,” *Psychological Medicine*, vol. 49, no. 9, pp. 1426–1448, 2019.

[9] T. Chen, S. Kornblith, M. Norouzi, and G. E. Hinton, “A simple framework for contrastive learning of visual representations,” *arXiv:2002.05709v1*, 2020.

[10] E. Bliss-Moreau, L. A. Williams, and A. C. Santistevan, “The immutability of valence and arousal in the foundation of emotion,” *Emotion*, 2019.

[11] J. J. Gross and R. F. Munoz, “Emotion regulation and mental health,” *Clinical Psychology*, vol. 2, no. 2, pp. 151–164, 1995.

[12] J. Li, L. Liu, T. D. Le, and J. Liu, “Accurate data-driven prediction does not mean high reproducibility,” *Nature Machine Intelligence*, 2020.

[13] A. Gelman, “Correlation does not even imply correlation,” *Statistical Modeling, Causal Inference, and Social Science (BLOG)*, 2014.

[14] B. Scholkopf, “Causality for machine learning,” *arXiv:1911.10500v1*, 2019.

[15] M. J. van der Laan and S. Rose, *Targeted Learning - Causal Inference for Observational and Experimental Data*. New York: Springer International, 2011.

[16] M. J. van der Laan and R. J. C. M. Starmans, “Entering the era of data science: targeted learning and the integration of statistics and computational data analysis,” *Advances in Statistics*, 2014.

[17] A. A. Aarts *et al.*, “Estimating the reproducibility of psychological science,” *Science*, vol. 349, no. 6251, 2015.

[18] D. Szucs and J. P. A. Ioannidis, “Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature,” *PLOS ONE Biology*, 2017.

[19] G. Parascandolo, N. Kilbertus, M. Rojas-Carulla, and B. Scholkopf, “Learning independent causal mechanisms,” *Proceedings of the 35 th International Conference on Machine Learning Learning*, 2018.

[20] M. Besserve, A. Mehrjou, R. Sun, and B. Scholkopf, "Counterfactuals uncover the modular structure of deep generative models," *arXiv:1812.03253v2*, 2019.

[21] R. Suter, D. Miladinovic, S. Bauer, and B. Scholkopf, "Interventional robustness of deep latent variable models," *arXiv:1811.00007v1*, 2018.

[22] F. Locatello, G. Abbati, T. Rainforth, T. Bauer, S. Bauer, B. Scholkopf, and O. Bachem, "On the fairness of disentangled representations," *arXiv:1905.13662v1*, 2019.

[23] C. Louizos, K. Swersky, Y. Li, M. Welling, and R. Zemel, "The variational fair autoencoder," *arXiv:1511.00830*, 2017.

[24] E. Creager, D. Madras, J.-H. Jacobsen, M. A. Weis, K. Swersky, T. Pitassi, and R. Zemel, "Flexibly fair representation learning by disentanglement," *arXiv:1906.02589v1*, 2019.

[25] H. Hosoya, "Group-based learning of disentangled representations with generalizability for novel contents," *Proc. 28th IJCAI*, 2019.

[26] D. Bouchacourt, R. Tomioka, and S. Nowozin, "Multi-level variational autoencoder: learning disentangled representations from grouped observations," *arXiv:1705.08841v1*, 2017.

[27] K. Bousmalis, G. Trigeorgis, N. Silberman, D. Krishnan, and D. Erhan, "Domain separation networks," *arXiv:1608.06019*, 2016.

[28] O. Press, T. Galatni, S. Benaim, and L. Wolf, "Emerging disentanglement in auto-encoder based unsupervised image content transfer," *ICLR*, 2019.

[29] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Transactions on pattern analysis and machine intelligence*, 2013.

[30] B. M. Lake, T. D. Ullman, J. B. Tenenbaum, and S. J. Gershman, "Building machines that learn and think like people," *Behavioral and Brain Sciences*, vol. 40, 2017.

[31] M. J. Vowels, N. C. Camgoz, and R. Bowden, "Gated variational autoencoders: Incorporating weak supervision to encourage disentanglement," *arXiv:1911.06443v1*, 2019.

[32] M. J. Vowels, R. Bowden, and N. Camgoz, "Deriving representations of facial expression," Master's thesis, Centre for Vision, Speech, and Signal Processing, University of Surrey, 2019.

[33] M. J. Vowels, N. C. Camgoz, and R. Bowden, "NestedVAE: Isolating common factors via weak supervision," *Conference on Computer Vision and Pattern Recognition*, 2020.

[34] K. Holstein, J. W. Vaughan, H. Daume III, M. Dudik, and H. Wallach, "Improving fairness in machine learning systems: what do industry practicioners need?," *arXiv:1812.05239v2*, 2019.

[35] R. Nabi, D. Malinsky, and I. Shpitser, "Optimal training of fair predictive models," *arXiv:1910.04109v1*, 2019.

[36] D. Slack, S. Hilgard, E. Jia, S. Singh, and H. Lakkaraju, "How can we fool LIME and SHAP? adversarial attacks on post hoc explanation methods," *arXiv:1911.02508v1*, 2019.

[37] P. Hilpert, T. R. Brick, C. Flueckiger, M. J. Vowels, E. Ceuleman, P. Kuppens, and L. Sels, "What can be learned from couple research: Examining emotional co-regulation processes in face-to-face interactions," *Journal of Counseling Psychology*, 2019.