Seeing the Forest from the Trees: A Novel Deep Learning-Driven Aggregate Embedding for Group-Level Analysis of Public Health Data

Alexander MacLean — Vision and Image Processing Group, University of Waterloo
Yang Yang — WHISTL Lab, University of Waterloo
Helen Chen — WHISTL Lab, University of Waterloo
Alexander Wong — Vision and Image Processing Group, University of Waterloo
Email: {alex.maclean, y24yang, helen.chen, a28wong}@uwaterloo.ca

Abstract

In the years since the COMPASS dataset initiative was begun, many important research questions have been investigated using its large amount of health information pertaining to high school students across Canada, with findings guiding many decisions made by policy makers [1]. However, to use traditional statistical methods, specific data points must be selected by researchers to include in the analysis, leading to possible unexpected relationships and connections across the study's 280 data points being missed. As well, most analysis is done on a per-student basis, while policies are often implemented at the school level, so understanding behaviours across a school's population can make it easier for school decision makers to interpret findings. Motivated by these goals, this study introduces a novel deep learning-driven aggregate embedding method to determine group-level representations for individual schools from student-level survey responses based on architecture introduced in Variational Autoencoders [2]. This study aims to produce a method which allows for new patterns to be identified in the COMPASS data and for the resulting embedded representations to be applied in future analysis.

1 Introduction

The COMPASS system [1] is an ongoing longitudinal study being conducted in Canadian high schools, with the goal being to monitor many facets of the health of students, notably obesity, tobacco use, other substance abuse, and bullying. From these data, COMPASS aims to better understand the contexts and environments that lead to both positive and negative health outcomes for the youth of today, allowing researchers to guide practices at the community level or higher which can improve said health outcomes.

There have been many studies published using data from COMPASS. [3] examined alcohol patterns in the years preceding and following Ontario allowing alcohol to be sold in grocery stores. [4] analyzed how bullying rates in youth affect future health measured by Body Mass Index (BMI). [5] studied how schools' disciplinary policies for cannabis offences are connected to reported use rates of cannabis. These examples show that the dataset being available has provided the opportunity for insights to be drawn; however, each of these studies makes use of only a small subset of data available to answer the desired research questions. COMPASS contains many data points, but researchers are generally selecting a few pieces of data relevant to their question - alcohol related questions for [3] or cannabis related questions for [5]. In doing so, much of the information contained in the rest of 280 data points is being discarded and unexpected patterns may be missed entirely.

One major goal for researchers is to be able to build transition models to predict if a certain school is going to move into a state of higher or lower risk in the future. There are a number of hurdles blocking progress on this task, the first of which is that most of the data are at the student-level, while ideally transition analysis would be done at the school-level. Additionally, researchers would like to avoid having to hand-select features for their specific research question, if it can be avoided. Motivated by this problem, the purpose of this study is to develop a method by which a representation of an individual school's health or risk can be found via an amalgamation of the entirety of the responses from the students at that school beyond simple statistical measures such average or variance of response values.

2 Methodology

2.1 COMPASS Dataset

The portion of COMPASS being use in this study is from 2018, and contains 280 data points measured from 74,501 students attending 136 schools from across Canada, in Ontario, Alberta, Quebec, and British Columbia. It contains demographic and personal data, such as ethnicity, primary language spoken, sex, height, and weight, as well as multiple choice type questions regarding the students' behaviour and feelings relating to health and wellness topics. As this dataset has been analyzed extensively, this study makes use of a pre-processed subset containing students with complete responses and which has been filtered to combine relevant points together (e.g. one question asks "If you do not eat breakfast every day, why do you skip breakfast?" with separate data points for each mutually exclusive option - these data points would be combined into a single feature in the processed dataset).

Included in this processed dataset are target features which have been identified as measures of health and risky behaviour in students, for which the discovery of patterns in other features have been deemed insightful and for which a transition model, to predict whether a school is likely to have students engage in said behaviours more or less, would be useful for policy makers to guide possible interventions. These targets include the rates of use of four substances: tobacco cigarettes, e-cigarettes, alcohol, and cannabis, with values ranging from never having used the substance to using the substance daily. This processed dataset contains 33 features.

2.2 Task

Using the COMPASS dataset, the task of this study is to determine a neural-network based architecture and processing pipeline which finds a representation at the school-level from the COMPASS responses of students in that school's population which can be used for pattern recognition and future transition analysis tasks.

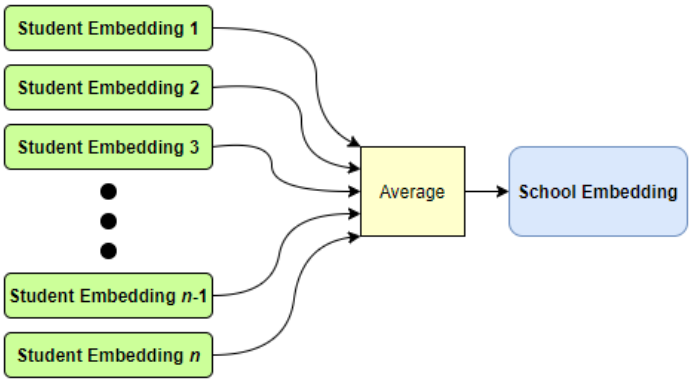

Fig. 1: Student-level embeddings are aggregated by averaging into a school-level embedding.

One challenge for combining the levels of data is that schools have varying numbers of students in their populations. Deep learning methods generally require consistent dimensionality in their input, but a proposed method would need to work correctly for populations of 100 or 1000 students. Towards this, the goal became to create school-level representations based on an aggregation of student-level representations. If a system for embedding student responses into a latent representational space is able to be done, then a combination of said embeddings should lead to a representation of the school which takes into account information from each

student. By choosing the aggregation to be done via averaging, the method will work on populations containing an arbitrary number of students as desired. An overview of this aggregation is shown in Figure 1. In order for this to work however, the student level representation requires that the embeddings be in a space where moving from point to point, or sampling between points, has intuitive and interpretable properties so that when averaged together, the resulting embedding is representative of the combination of students from which it is created.

2.3 Variational Architecture

To this end, an architecture inspired by Variational Autoencoders (VAEs) was chosen [2]. VAEs are a modification of the Autoencoder (AE), which is a deep learning method which embeds a given input into a (generally) lower dimensional space by learning to reconstruct the input, but while reducing the dimensionality internally. Thus, the model should have to discard information less relevant for reconstruction, meaning the reduced representation, or latent space representation, is more dense and compact. VAEs modify AE architecture by introducing variability into the layer representing the latent space. Instead of the series of weights and nodes leading to deterministic embeddings, the layer leading to the embedding learns statistics for the distribution of the latent space for the given sample by determining values for the mean and variance of that given latent variable. A value is then sampled from that distribution to be passed on to the following layers as desired. Since the calculated embeddings for samples are now stochastic in nature, the model tends to learn to be more robust to fluctuations in the input, as well as providing a latent space across which movement between samples can be done intuitively.

Additionally, and importantly for the purposes of this study, since the model has to be robust to variation in the latent space, VAEs are successful at learning complex, non-linear embeddings tailored to the data at hand. Since the desired task is not just dimensionality reduction, the entirety of the VAE architecture was not used. Instead, since it is desired that the school, and by extension student, representations contain information pertaining to the target variables, they are included in the development of the embeddings. The architecture was structured such that the model learned to predict each of the four targets, with the final layer before classification being the distribution-sampling layer seen in VAEs, rather than learning to reconstruct the input as in a standard VAE. For the purposes of this study, this architecture with the variational structure included is called the V-Classifier.

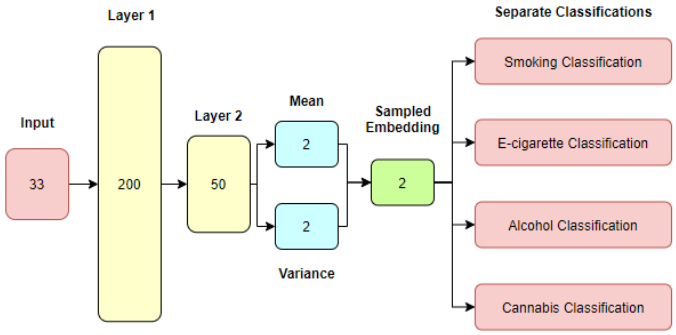

Fig. 2: V-Classifier Architecture. The values as the Sampled Embedding layer are taken as the embedding for a given student's input, which are used in Figure 1 as Student Embedding.

A diagram of the chosen architecture is shown in Figure 2. The latent space was chosen to have dimension 2 to restrict information as much as possible and to make visualization of the resulting latent space possible without further dimensionality reduction. The model was trained using using Keras for 20 epochs with the RMSprop optimizer.

3 Results and Discussion

3.1 Student-Level Embedding

Figure 3 displays the learned embeddings for individual students. Points are coloured based on the responses from the students to each of the four substance use questions.

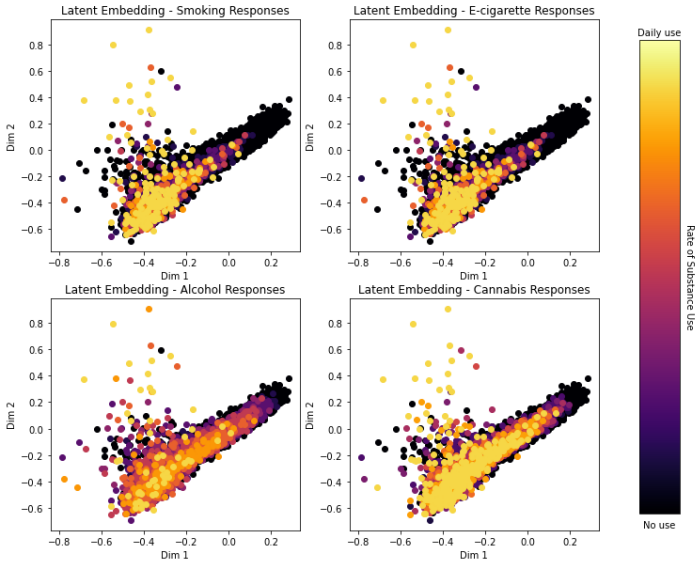

Fig. 3: Student-level latent space from V-Classifier. In each plot, the colours represent the responses for individuals pertaining to specific substances. Black represents students who reported no substance use, with brighter colours representing higher rates of use up, to daily use.

Note that students with higher rates of use of each substance tend to be focused in the bottom left of the distribution of embeddings, suggesting that the model is able to identify patterns in the student features which are able to be used to distinguish individuals more likely to use substances at higher rates. This result is promising for the embeddings to be useful in creating interpretable representations at the school level, since the input from each student's embedding should carry information to be applied to the school's representation.

Additionally, at least with labels corresponding to low rates of use, alcohol and cannabis use is present in the entire distribution of students while with cigarettes and e-cigarettes, there are areas of the embedding space where essentially all students responded that they had never used the substance (coloured black). This suggests that the former substances are more widely accepted and used by students at least at low levels.

3.2 School-Level Representation

As desired, the V-Classifier architecture allows for moving from student-level embeddings to school-level representations by averaging the responses from students at each of the 130 schools. Figure 4 displays the resulting school-level representation. Points are coloured based on the percentage of students whose responses indicated that they use a particular substance at a rate of at least once a week.

For each substance, there is a similar pattern across schools, where schools located in the lower left have higher substance use rates and those in the upper right have lower use. There is generally a gradient when moving from high- to low-use, but particularly with alcohol the high-use schools are somewhat more centered in the distribution.

Despite this pattern, there are some outliers. In particular, the school closest to the bottom left corner has low rates for each substance. When analyzing Figure 3, even though higher rates are skewed towards that direction, there are low usage student responses still in that region. It is interesting that whatever patterns the model is using to identify students likely to use substances at a higher rate are seen in that school even though its substance use rates are lower. Further exploration is needed to examine this case to see what led to this unique embedding.

Compared to Figure 3, where many students reported no use of e-cigarettes but at least limited use of alcohol and/or cannabis, the school-wide plots of those three substances look very similar. This suggests that, while many students will not use e-cigarettes, those that do are more likely to use that substance on a more regular basis (at least weekly) when compared to rates of use for alcohol or cannabis.

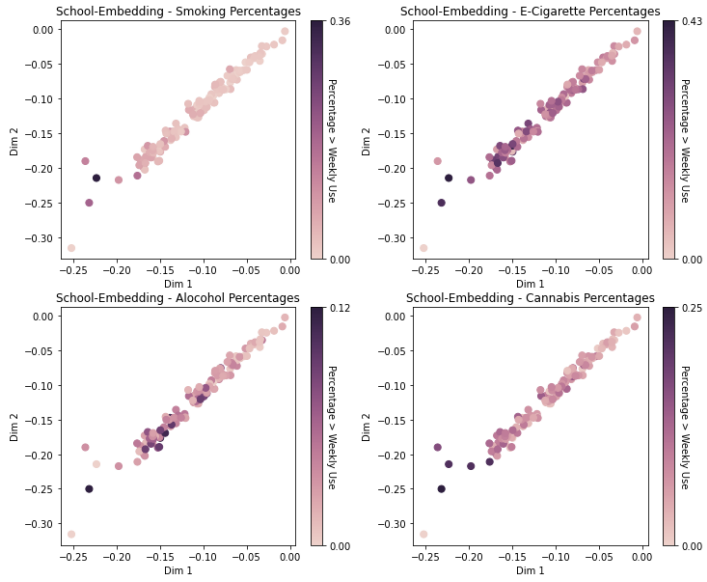

Fig. 4: Embeddings at the school-level. The student level embeddings were averaged based on school-ID, and the colour gradient in each plot represents the percentage of students who responded that they used the given substance at a rate of at least once a week. Darker colours represent higher usage percentages.

In any case, this representation does appear to capture school-wide patterns about substance use, suggesting that future work can be done to further identify patterns and gain insights to better understand the data, as well as to use such a model to build systems to predict future behaviours and guide school- and population-wide policies.

3.3 Recommendations

The proposed deep learning-driven aggregation is one method for merging information at the school- and student-level. However, there exists in the COMPASS dataset information that is only at the school-level from surveys sent to representatives at each school. In this study, the only learning that occurred is at the student-level; that being the use of the V-Classifier to learn individual embeddings which are then aggregated to form the group-level representations. Including information that exists only at the group-level would require some modification to the architecture, either by introducing learning after aggregation or by imbuing each student's data with information from their school. Further work is need to determine which methods would best incorporate the additional data.

As mentioned previously, one main goal of determining representations for schools from this dataset is to be able to make predictions about whether a school will move into a higher- or lower-risk grouping in coming years. There are a subset of schools for which there is data across two or three years with which such analysis can be done. Future work can then investigate whether the methods introduced in this study are able to lead to successful transition analysis, either by using the embedded representations provided by the V-Classifier or possibly by building a further transition classification model or Recurrent Neural Network model which incorporates the variational architecture described here.

4 Conclusion

The results of this study show that aggregation of student-level embeddings found via a deep learning model leads to a group-level representation which is consistent with desired target features relating to students' substance use. By expanding on this work, future studies will aim to identify previously unseen patterns and produce tools to predict relevant changes in health of student populations. Hopefully such findings will continue to use the immense work done to curate the COMPASS dataset to guide the development of public health policies to create a healthier environment for youth across Canada.

Acknowledgments

The COMPASS study has been supported by a bridge grant from the CIHR Institute of Nutrition, Metabolism and Diabetes (INMD) through the "Obesity – Interventions to Prevent or Treat" priority funding awards (OOP-110788; awarded to SL), an operating grant from the CIHR Institute of Population and Public Health (IPPH) (MOP-114875; awarded to SL), a CIHR project grant (PJT-148562; awarded to SL), a CIHR bridge grant (PJT-149092; awarded to KP/SL), a CIHR project grant (PJT-159693; awarded to KP), and by a research funding arrangement with Health Canada (#1617-HQ-000012; contract awarded to SL).

References

[1] S. T. Leatherdale, K. S. Brown, V. Carson, R. A. Childs, J. A. Dubin, S. J. Elliott, G. Faulkner, D. Hammond, S. Manske, C. M. Sabiston, R. E. Laxer, C. Bredin, and A. Thompson-Haile, "The COMPASS study: a longitudinal hierarchical research platform for evaluating natural experiments related to changes in school-level programs, policies and built environment resources," *BMC Public Health*, vol. 14, no. 1, Apr. 2014. [Online]. Available: https://doi.org/10.1186/1471-2458-14-331

[2] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," 2013.

[3] M. R. Gohari, R. J. Cook, J. A. Dubin, and S. T. Leatherdale, "The impact of an alcohol policy change on developmental trajectories of youth alcohol use: examination of a natural experiment in canada," *Canadian Journal of Public Health*, Aug. 2020. [Online]. Available: https://doi.org/10.17269/s41997-020-00366-7

[4] N. Hammami, A. Chaurasia, P. Bigelow, and S. T. Leatherdale, "Exploring gender differences in the longitudinal association between bullying and risk behaviours with body mass index among COMPASS youth in canada," *Preventive Medicine*, vol. 139, p. 106188, Oct. 2020. [Online]. Available: https://doi.org/10.1016/j.ypmed.2020.106188

[5] M. Magier, K. A. Patte, K. Battista, A. G. Cole, and S. T. Leatherdale, "Are school substance use policy violation disciplinary consequences associated with student engagement in cannabis?" *International Journal of Environmental Research and Public Health*, vol. 17, no. 15, p. 5549, Jul. 2020. [Online]. Available: https://doi.org/10.3390/ijerph17155549