2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

Fusing multimodal neuroimaging data with a variational autoencoder

Eloy Geenjaar[1,2], Noah Lewis[1], Zening Fu[1], Rohan Venkatdas[1,3], Sergey Plis[1], Vince Calhoun[1]

Abstract— **Neuroimaging studies often collect multimodal data. These modalities contain both shared and mutually exclusive information about the brain. This work aims to find a scalable and interpretable method to fuse the information of multiple neuroimaging modalities into a lower-dimensional latent space using a variational autoencoder (VAE). To assess whether the encoder-decoder pair retains meaningful information, this work evaluates the representations using a schizophrenia classification task. The linear classifier, trained on the representations obtained through dimensionality reduction, achieves an area under the curve of the receiver operating characteristic (ROC-AUC) of 0.8609. Thus, training on a multimodal dataset with functional brain networks and a structural magnetic resonance imaging (sMRI) scan, leads to dimensionality reduction that retains meaningful information. The proposed dimensionality reduction outperforms both early and late fusion principal component analysis on the classification task. -**

Clinical relevance – **This work examines the interplay between neuroimaging modalities and their relation to mental disorders. This allows for more complex and rigorous analysis of multimodal neuroimaging data throughout clinical settings.**

I. INTRODUCTION

Multimodal neuroimaging data is abundantly available and although approaches that seek to combine these data, e.g., JointICA [1], and more recently multimodal subspace analysis [2] focus on linear decompositions, recent work on multimodal deep learning offers the benefits of additional flexibility which can also capture nonlinear relationships. Multimodal deep learning research mostly focuses on the relationship between audio, images, and/or text [3]. The exciting new direction of multimodal representation learning, together with growing evidence that deep learning representations can provide robust biomarkers [4], paves the way for multimodal representation learning in neuroimaging.

Fusing modalities into lower-dimensional representations can lead to biomarkers that more robustly predict changes associated with mental illnesses [5]. An important downside to deep learning techniques is that their non-linear nature can present challenges to interpretation, which undermines their applicability to medical problems. Interpretability is, therefore, an important consideration in this work.

Recent work in multimodal deep learning applied to neuroimaging has focused on information maximization between representations extracted from two modalities [6], [7] or by translating between modalities [8]. This work aims to learn a continuous manifold of multiple modalities so that they are represented in a locally Euclidean space. The model architecture that is used is a variational autoencoder (VAE) [9], which maximizes a lower bound on the log-likelihood of the data's marginal distribution. Other work on multimodal VAEs focuses on a factorization of shared and private subspaces [10] and uses a separate encoder for each modality. In this work, we intentionally force all of the modalities to populate the same shared subspace by using a single encoder-decoder pair for all modalities. This, for example, allows for natural interpolation between modalities, similar to the interpolation between different digits in the MNIST dataset [9].

To provide an initial assessment of the representations extracted by the VAE through unsupervised training, they are evaluated with a schizophrenia classification task. The method is compared to both early and late fusion principal component analysis (PCA). Schizophrenia is a mental illness that is characterized by complex interconnected changes in dynamics and functional connectivity. To understand how the brains of patients with schizophrenia differ from controls it is imperative to piece together information from multiple modalities [5]. In this work, we treat a structural MRI (sMRI) volume and each of the intrinsic functional brain networks that are extracted from resting-state functional MRI (rs-fMRI) data using NeuroMark [11] as separate modalities. The multimodal terminology also stems from the use of functional modes for intrinsic functional brain networks.

An important consideration when choosing our method is that a VAE is a generative model. It can therefore decode locations in the latent space back to brain space and either generate new data or help interpret locations in the latent space. The regions that have previously been linked to schizophrenia include the thalamus, cerebellum, caudate, superior temporal gyrus, most of the visual system (e.g., lingual gyrus, occipital gyrus [12]), and the supplementary motor area [13].

II. CONTRIBUTIONS

This work introduces a generative and interpretable approach for fusing multiple neuroimaging modalities with the following properties:

- Scaling in the number of parameters is $\mathcal{O}(1)$ with the number of modalities.
- The model can generate new samples outside of the training distribution, this could be used to perform data augmentation or interpret individual and group differences.

*This work was not supported by any organization

[1] Tri-institutional Center for Translational Research in Neuroimaging and Data Science (TReNDS), Georgia State, Georgia Tech, Emory, Atlanta, GA 30303, USA e.geenjaar@gsu.edu

[2] Faculty of Electrical Engineering, Mathematics & Computer Science, TU Delft, Delft, the Netherlands

[3] Lambert High School, Suwanee, GA, USA

978-1-7281-1178-0/21/$31.00 ©2021 IEEE

- The model extracts representations for each modality separately but does so in a single model. Although the representations themselves are not made up of each modality, the weights that are used to extract those representations are shared. The method thus combines advantages from both early fusion and late fusion. The shared weights of early fusion, and the separate representations for each modality as in late fusion.

III. Method

A. Problem setting

Let $\{M_i = \{x_{i,j}, ..., x_{i,N}\}\}_{i=1,...,n}$ be a set of modalities with N subjects and n modalities. Instead of learning each modality with a separate decoder, we enforce a shared subspace. Further, to make this approach scalable to a large number of modalities and because we already use a shared decoder, we also only use one encoder for all modalities. This forces the features that are learned for each modality to be similar and makes sure the model scales well in terms of memory usage. Further, given that neuroimaging datasets are considered small compared to more commonly used deep learning datasets, using multiple encoders may lead to overfitting. The encoder-decoder couple is optimized for the log-likelihood of the marginal distribution of M and each volume is treated as an independent sample. The integral over the marginal distribution of M, $p_\theta(x) = \int p_\theta(z)p_\theta(x|z)dz$ is intractable. We, therefore, optimize the evidence lower bound (ELBO) [9]. The variational autoencoder (VAE) consists of an encoder and a decoder. The encoder $q_\phi(z_{i,j}|x_{i,j})$, parameterized as a convolutional neural network (CNN) with parameters ϕ, estimates latent variable $z_{i,j}$ given a data point $x_{i,j}$. The decoder $p_\theta(x_{i,j}|z_{i,j})$, also parameterized as a CNN, but with parameters θ, reconstructs the original sample $x_{i,j}$ from the estimated latent variable $z_{i,j}$. The ELBO is made up of a log-likelihood maximization of reconstructions and a KL-divergence minimization between a prior of our choosing $p_\theta(z)$ and the posterior approximated by the encoder. The prior in this work is a diagonal multivariate Gaussian centered at 0, with a standard deviation of 1. The approximate posterior that is sampled from to obtain $z_{i,j}$ is also a multivariate Gaussian, parameterized by the mean $\mu_{i,j}$ and variance $\sigma_{i,j}$ predicted by the encoder.

B. Classification

To evaluate whether the dimensionality reduction retains meaningful information, we set up a classification task. The model is first trained using 10-fold cross-validation, where each fold of subjects is used as a test set once and the other 9 folds are used to train on. The validation set is randomly selected as a stratified 10% of the subjects in the training set. After training the VAE, the weights in the VAE are frozen. The complete dataset is then embedded using the encoder $q_\phi(z_{i,j}|x_{i,j})$, where instead of sampling $z_{i,j}$ from its estimated multivariate Gaussian, we use the estimated mean μ as our latent variable $z_{i,j}$. This is to make sure there is no stochasticity in the inference process, furthermore using both the mean $\mu_{i,j}$ and the variance $\sigma_{i,j}$ as features did not improve the model over only using $\mu_{i,j}$ in our preliminary results. The representations of the training and validation sets are stacked and used as input for a machine learning model.

The VAE is compared to both early and late fusion PCA, the first uses a separate decomposition for each modality and concatenates those as features for the classifier. The latter uses a single decomposition on all of the modalities at once. This results in a smaller number of features per subject because late fusion and the VAE decompose each modality separately and concatenates those features. The maximum number of components for the PCA decomposition is limited to the number of subjects in the training and validation set, which means the results for 1024 latent dimensions are not available for the PCA decompositions.

The estimated latent variable $z_{i,j}$ is a low-dimensional representation of a volume with a dimensionality l. Given that each subject has n different modalities, each subject j will also have n representations $z_{i,j}, ..., z_{n,j}$. These representations can be concatenated for a subject to create a feature vector with a size of $n \times l$. The subject-by-feature matrix can be used as input for a classifier. In this case, we train a support vector machine (SVM) to predict whether subjects in a held-out test set are patients with schizophrenia. Given that each modality is represented using l features, we can extract the feature importance for all nl features and then sum the features for each modality, to get feature importance for each of the n modalities. The feature importance helps us understand how brain changes related to schizophrenia are jointly represented in multiple modalities.

The classification task is evaluated using the area under the curve (AUC) of the receiver operator characteristic (ROC). We evaluate the model for 5 different seeds to ensure robustness, these experiments are performed with a latent dimensionality of 128. To evaluate the effect of the number of latent dimensions on the information retained in the representations, we set the seed to 42 and train the model with four different latent dimensionalities 128, 256, 512, 1024. The performance is determined by averaging the ROC-AUC over the 10 training folds. The encoder and decoder trained on the first fold are used to create the figures and to determine the feature importance for each modality.

C. Data

The datasets used in this study are FBIRN, B-SNIP, and COBRE, each dataset, and modality was processed using NeuroMark [11] to obtain 53 independent component networks (ICNs). These 53 ICNs, together with a structural MRI scan for each subject are considered to be separate modalities, so n=54. The sMRI data is preprocessed using SPM 12 in a Matlab 2016 environment. The data is then segmented into modulated gray matter volumes (GMV) and smoothed with a 6mm FWHM Gaussian kernel. Each ICN is a 53-by-63-by-52 volume, the sMRI volumes are resized to the same size using Scipy [14]. The values in each volume are rescaled to [-1, 1] by dividing the values in a volume by their maximum, which is also sometimes referred to as maximum absolute scaling. The dataloader and transformations were implemented with the help of TorchIO [15].

D. Implementation

The batches are constructed by loading the 53 ICN volumes and an sMRI volume for a subject and concatenating them into a batch. The volumes are loaded per subject to minimize disk access. The ICNs for a subject are all saved in one file, so loading them into a batch together leads to a smaller number of disk accesses and reduces training time. The loss calculated over a batch of subjects is therefore balanced across modalities.

The code for the VAE and PCA models, inference, and training are implemented using PyTorch [16], Catalyst [17], and NumPy [18]. The VAE uses a convolutional encoder and decoder pair, each of the layers uses a 3-voxel kernel, a stride of 2, and 1-voxel padding. The channel sizes in the encoder are $1 \rightarrow 64, 64 \rightarrow 128, 128 \rightarrow 256, 256 \rightarrow 512$ and $512 \rightarrow 256, 256 \rightarrow 128, 128 \rightarrow 64, 64 \rightarrow 32, 32 \rightarrow 16, 16 \rightarrow 1$ in the decoder, the last layer in the decoder uses stride 1 and no bias parameters. The activation function after each convolutional layer is a ReLU [19], except for the last layer in the decoder, which uses a hyperbolic tangent function to map the output between [-1, 1] to match the input range. The last convolutional layer in the encoder produces an output with shape: 4-by-4-by-4 and 256 channels, this output is flattened and mapped to the mean $\mu_{i,j}$ and variance $\sigma_{i,j}$, which are used to construct a diagonal multivariate Gaussian from which $z_{i,j}$ is sampled. To make sure the VAE is fully differentiable, we train the model using the reparameterization trick [9]. The classification evaluations in the latent space are implemented using RAPIDS AI [20] to make sure highly parallelizable computations are performed on the GPU and to minimize costly CPU→GPU and GPU→CPU data transfers. The experiments were performed on an NVIDIA DGX-1 V100.

E. Latent structure

Most of the modalities in this paper are intrinsic networks, which are obtained through independent component analysis (ICA). The independence in the spatial volumes for those components leads to a latent space that clusters modalities, which is shown in Figure 1. The plot depicts a t-SNE [21] projection, that can visualize the distances between points in the VAE's 512-dimensional latent space in 2D. Interestingly, the ICNs that belong to the same domain are generally clustered together, such as ICNs in the cerebellum. It is also clear from Figure 1 that the sMRI cluster is located relatively far away from the other modalities in the latent space. The ICNs represent localized functional brain networks, whereas the sMRI volume represents all of the structures in the brain. There is more inter-subject variance to be modeled for the sMRI volumes than for the spatially localized ICNs. This likely contributes to the sMRI cluster being further away from the latent ICN clusters.

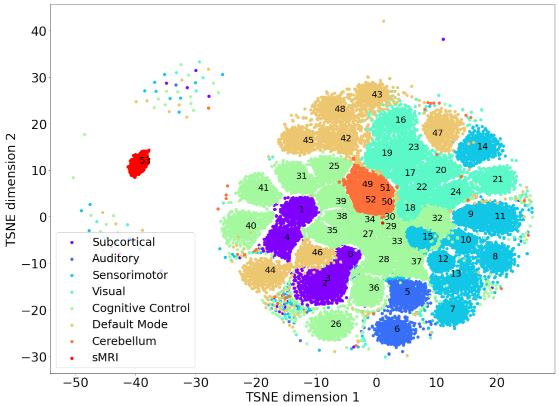

Fig. 1. A 2D t-SNE projection of the VAE's 512-dimensional latent space, each number indicates a different modality starting at 0. Each color indicates the domain that the intrinsic networks belong to. Each subject is represented by 54 points in this plot, one for each modality.

IV. Results

A. Classification

The average ROC-AUC for the five models trained with different seeds and a latent dimensionality of 128 is 0.8374, with a standard deviation of 0.0027. This shows that the model robustly learns a latent space, where patients with schizophrenia and controls are linearly separable.

TABLE I
THE ROC-AUC RESULTS ACROSS MULTIPLE METHODS AND LATENT DIMENSIONS. THE VAE OUTPERFORMS THE OTHER METHODS.

Latent dimension	128	256	512	1024
VAE	**0.8353**	**0.8569**	**0.8609**	**0.8539**
Early fusion PCA	0.8229	0.8358	0.8302	X
Late fusion PCA	0.5012	0.5078	0.4914	X

The results in Table I show that the VAE outperforms both the early and late fusion PCA. Furthermore, increasing the latent dimensionality increases the meaningful information in the latent space up to 512 dimensions.

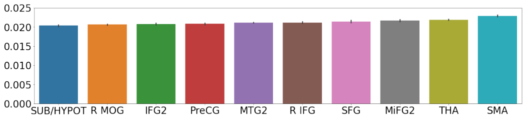

Fig. 2. The top 10 most important modalities, with their names on the x-axis and the importance that the SVM assigns to them (that sums to 1) on the y-axis. The plot shows the standard deviation over each of the 10 test folds as a vertical line for each modality.

The feature importances are calculated using the best model (seed=42 and 512 latent dimensions). The 10 modalities with the highest feature importance are shown in Figure 2, where the rightmost modality is the most important and the leftmost modality is the 10th most important. sMRI is the least important modality. The number of different modalities combined with the prior, that pulls the distributions of the modalities towards zero-mean unit-norm, limits the variance that can be modeled to represent the sMRI volumes. The variations that are modeled for sMRI in the latent space do not help linearly separate patients from controls.

The group differences that the VAE has learned can be interpreted by visualizing their latent group centers. These latent group centers are the average latent location of subjects within that group. The latent center for subjects diagnosed with schizophrenia can then be decoded and subtracted from the decoded latent center for healthy controls to show the group differences. The differences of the top five most important features are calculated, then thresholded at the 99th quantile highest values for each modality, and then summed to create Figure 3. The figure compares the learned differences with the voxelwise differences of the spatial ICA components that correspond to the five most important modalities. The results are remarkably similar, which shows that important group differences are retained even after a large dimensionality reduction.

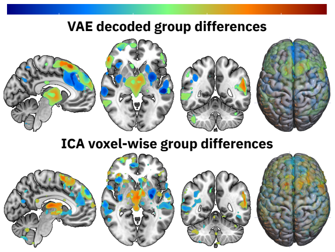
VAE decoded group differences
ICA voxel-wise group differences

Fig. 3. The combined differences between patients diagnosed with schizophrenia and healthy controls for the five modalities with the highest feature importance.

V. Conclusion

When the number of modalities increases for multimodal learning, it may not be feasible or optimal to learn a separate encoder-decoder pair for each modality. This is especially true for small datasets where overfitting due to overparameterization is a potential problem. This work takes the approach of joint multimodal representation learning by modeling the marginal distribution of all the modalities together. The VAE learns subspaces in the latent space that can linearly separate healthy controls from subjects diagnosed with schizophrenia. The proposed framework is easy to generalize to more modalities, although modalities like functional connectivity will require some engineering because the network currently expects each modality to be a 53x52x63 volume.

VI. Future work

The independence of spatial ICA components is reflected in the latent space of our model, which leads us to believe that unprocessed volumes may be an important direction for fusing modality representations. Another way to tackle this problem is to enforce additional losses in the latent space or create an inductive bias in the architecture of the model. Furthermore, computing joint features (early fusion) and using those as inputs for the model may also increase multimodal fusion in the latent space.

References

[1] Calhoun, V. D., Adali, T., Pearlson, G. D., & Kiehl, K. A. (2006). Neuronal chronometry of target detection: fusion of hemodynamic and event-related potential data. Neuroimage, 30(2), 544-553.

[2] Silva, R. F., Plis, S. M., Adalı, T., Pattichis, M. S., & Calhoun, V. D. (2020). Multidataset Independent Subspace Analysis with Application to Multimodal Fusion. IEEE Transactions on Image Processing, 30, 588-602.

[3] Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., & Ng, A. Y. (2011, January). Multimodal deep learning. In ICML.

[4] Abrol, A., Fu, Z., Salman, M., Silva, R., Du, Y., Plis, S., & Calhoun, V. (2021). Deep learning encodes robust discriminative neuroimaging representations to outperform standard machine learning. Nature communications, 12(1), 1-17.

[5] Calhoun, V. D., & Sui, J. (2016). Multimodal fusion of brain imaging data: a key to finding the missing link (s) in complex mental illness. Biological psychiatry: cognitive neuroscience and neuroimaging, 1(3), 230-244.

[6] Fedorov, A., Geenjaar, E., Wu, L., DeRamus, T. P., Calhoun, V. D., & Plis, S. M. (2021). Tasting the cake: evaluating self-supervised generalization on out-of-distribution multimodal MRI data. arXiv preprint arXiv:2103.15914.

[7] Fedorov, A., Sylvain, T., Geenjaar, E., Luck, M., Wu, L., DeRamus, T. P., ... & Plis, S. M. (2020). Self-Supervised Multimodal Domino: in Search of Biomarkers for Alzheimer's Disease. arXiv preprint arXiv:2012.13623.

[8] Plis, S. M., Amin, M. F., Chekroud, A., Hjelm, D., Damaraju, E., Lee, H. J., ... & Calhoun, V. D. (2018). Reading the (functional) writing on the (structural) wall: Multimodal fusion of brain structure and function via a deep neural network based translation approach reveals novel impairments in schizophrenia. NeuroImage, 181, 734-747.

[9] Kingma, D. P., & Welling, M. (2013). Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114.

[10] Shi, Y., Siddharth, N., Paige, B., & Torr, P. H. (2019). Variational mixture-of-experts autoencoders for multi-modal deep generative models. arXiv preprint arXiv:1911.03393.

[11] Du, Y., Fu, Z., Sui, J., Gao, S., Xing, Y., Lin, D., ... & Alzheimer's Disease Neuroimaging Initiative. (2019). NeuroMark: an adaptive independent component analysis framework for estimating reproducible and comparable fMRI biomarkers among brain disorders. MedRxiv, 19008631.

[12] Salman, M. S., Du, Y., Lin, D., Fu, Z., Fedorov, A., Damaraju, E., ... & Calhoun, V. D. (2019). Group ICA for identifying biomarkers in schizophrenia:'Adaptive'networks via spatially constrained ICA show more sensitivity to group differences than spatio-temporal regression. NeuroImage: Clinical, 22, 101747.

[13] Schröder, J., Wenz, F., Schad, L. R., Baudendistel, K., & Knopp, M. V. (1995). Sensorimotor cortex and supplementary motor area changes in schizophrenia. The British Journal of Psychiatry, 167(2), 197-201.

[14] Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., ... & van Mulbregt, P. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. Nature methods, 17(3), 261-272.

[15] Pérez-García, F., Sparks, R., & Ourselin, S. (2020). TorchIO: a Python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning. arXiv preprint arXiv:2003.04696.

[16] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., ... & Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. arXiv preprint arXiv:1912.01703.

[17] Kolesnikov, S. (2018). Accelerated deep learning R&D. GitHub.

[18] Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., ... & Oliphant, T. E. (2020). Array programming with NumPy. Nature, 585(7825), 357-362.

[19] Glorot, X., Bordes, A., & Bengio, Y. (2011, June). Deep sparse rectifier neural networks. In Proceedings of the fourteenth international conference on artificial intelligence and statistics (pp. 315-323). JMLR Workshop and Conference Proceedings.

[20] Team, R. D. (2018). RAPIDS: Collection of Libraries for End to End GPU Data Science. NVIDIA, Santa Clara, CA, USA.

[21] Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. Journal of machine learning research, 9(11).