Modeling pain in the brain with conditional variational autoencoder

Sungwoo Lee (sungwoo320@gmail.com)
Jihoon Han (hahnz@g.skku.edu)
Choong-Wan Woo (waniwoo@g.skku.edu)
Center for Neuroscience Imaging Research, Institute for Basic Science, Suwon, South Korea
Department of Biomedical Engineering, Sungkyunkwan University, Suwon, South Korea
Department of Intelligent Precision Healthcare Convergence, Sungkyunkwan University, Suwon, Korea

Abstract:

Variational autoencoders (VAE) have received significant attention from the deep learning and neuroscience communities as the VAE is able to find meaningful low-dimensional representations from high-dimensional data. Although the VAE was originally developed as an unsupervised learning algorithm, recent advances in the VAE allow us to use it in a semi-supervised manner (e.g., conditional VAE). These new advances allow neuroscientists to condition the VAE on their experimental variables and identify the condition-free low dimensional representations. Here, we trained conditional VAEs to model brain responses to different levels of noxious heat stimulation with a functional Magnetic Resonance Imaging (fMRI) dataset (total *N* = 124). By conditioning the data on different levels of heat intensity, we extracted the condition-free low-dimensional latent variables with training and validation data (*n* = 87). Then, we were able to generate brain responses for any given conditions (i.e., heat intensity) from no-pain data in a test dataset (*n* = 37). Further analyses revealed that the condition-free latent variables can identify different individuals with high accuracy, suggesting that the latent variables contain idiosyncratic fMRI features of each individual. Overall, we show that the conditional VAE can model the effects of heat intensity as well as individual variability, providing a powerful analysis strategy both for population-level and personalized pain neuroimaging.

Keywords: variational autoencoder, pain, fMRI

Conditional VAE for fMRI

Finding meaningful low-dimensional representations of neural population activity is one of the major goals in neuroscience. In functional Magnetic Resonance Imaging (fMRI) studies, linear dimensionality reduction methods, such as PCA and ICA, have been the most popular strategies to find the low-dimensional neural representation. However, these methods depend only on linear relationships among variables and cannot leverage experimental condition information. Recent advances in deep learning have created a new opportunity to identify low-dimensional representations using non-linear generative models. In particular, a variational auto-encoder (VAE) is currently receiving significant attention from multiple fields (Kingma & Welling, 2013). Mathematically, the VAE can be viewed as a non-linear ICA or a probabilistic PCA when it uses a linear activation function (Khemakhem et al., 2020; Lucas et al., 2019; Rolinek et al., 2019; Zietlow et al., 2021). More importantly, the VAE can also be trained in a semi-supervised manner (Kingma & Mohamed, 2014). For example, the conditional VAE (cVAE) can incorporate condition information into the training of generative models (Sohn et al., 2015). The cVAE does not only allow us to identify condition-free low dimensional representations of neural activity, but also generates and predicts neural activity for different conditions (Lim et al., 2018). Overall, the cVAE has the potential to identify meaningful low-dimensional representations of brain activity in response to different levels of painful stimulation.

Methods

Pain fMRI dataset. We collected an fMRI dataset (total *N* = 124) with a simple pain task design, which consisted of a pre-stimulus period (3~5s), pain stimulation (12s), and pain rating (5s) within one trial (**Fig. 1a**). We used six different levels of heat stimulus intensity (45.0 to 47.5°C with 0.5°C increment). For each participant, we delivered heat stimulation 96 times (i.e., 96 trials). We divided the data into training (*n* = 79), validation (*n* = 8), and test sets (*n* = 37). All results in **Fig. 2** are the results of test sets.

Conditional VAE. The cVAE consisted of encoder and decoder parts. In the training step, the condition variable (here, heat stimulus intensity) was used to train the encoder and to extract condition-free latent representations from fMRI data. This condition variable was used in the decoder part to generate the condition-dependent data, which is the original fMRI data (**Fig. 1b**). Then, the trained cVAE can be used to generate new brain maps from unseen data and different condition variables, allowing us to test whether the cVAE model successfully extracts the condition-free low dimensional features from the data. Lastly, we can further analyze the model to better understand the low-dimensional latent space.

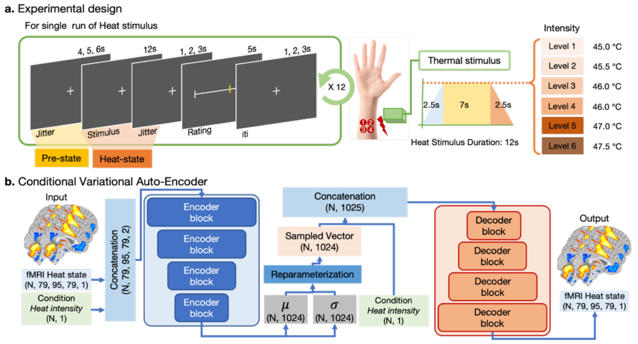

Fig. 1. (a) Experiment design and **(b)** conditional VAE model structure.

Results

Training and testing cVAE model: As shown in **Fig. 2a**, we first obtained the group-level contrast map for the heat level 6 versus 1 using the raw fMRI data. For thresholding, we used the false discovery rate correction $q < 0.05$ for multiple comparisons. In addition, we conducted the term-based decoding with a large-scale meta-analysis database, Neurosynth, to examine whether the results are sensible, which is shown as a wordcloud in **Fig. 2a**.

Then, we obtained the group-level contrast map for the heat levels 6 versus 1 based on the cVAE generated fMRI data using the actual fMRI data of levels 6 and 1 as inputs (**Fig 2b**).

Lastly, we obtained the group-level contrast map for the heat levels 6 versus 1 based on the cVAE generated fMRI data using the fMRI data from the pre-state period (i.e., no-pain condition) as inputs (**Fig. 2c**). All three conditions generated sensible Neurosynth decoding results, e.g., sensorimotor-related terms and also pain.

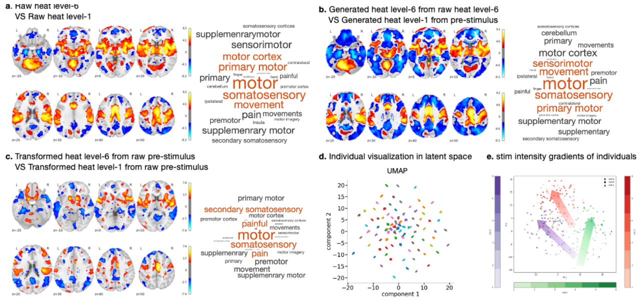

Figure 2. Group-level contrast maps. (a-c) The contrast maps of three conditions with FDR correction ($q < 0.05$). Color bars indicate t-values. The wordcloud shows the Neurosynth decoding results of each map. The sizes of the word represent the relative sizes of correlation values. The top five words are in red. All results are from the test dataset (n = 37). **(d)** The UMAP visualization of test data (n = 37) on the latent space (1024 dimensions). Different colors indicate different individuals. **(e)** The PCA visualization of the data on the latent space of VAE (1024 dimensions) trained without condition. Different colors indicate different individuals, and the color gradient represents stimulus intensities.

Analysis of low-dimensional latent space: We further analyzed the low dimensional representations learned by the cVAE model and found clustering of individuals when we visualized the data with the Uniform Manifold Approximation and Projection (UMAP) (**Fig. 2d**). This result is consistent with a previous study (Kim et al., 2021), which reported that the VAE is better at the individual identification compared to PCA or ICA.

VAE model without conditions: To further understand the low-dimensional latent variables from the VAE models, we trained an additional VAE without conditions and compared them to the cVAE results. Similar to the cVAE results, the VAE also showed the individual clustering, but within each individual cluster, we were able to observe the stimulus intensity information (**Fig. 2e**).

Conclusion

In this study, we successfully applied the cVAE to obtain condition-free low dimensional latent representations of brain response to painful heat. The cVAE was able to generate similar brain responses for any given conditions for unseen data. However, the condition-free representations still contained individuals' idiosyncratic features. This suggests the possibility of using cVAE to model both the effects of heat intensity and individual's unique features. Overall, this study provides a powerful data analysis strategy both for population-level and personalized pain neuroimaging.

Acknowledgments

This work was supported by IBS-R015-D1 (Institute for Basic Science; to C.-W.W.), 2019R1C1C1004512, 2021M3E5D2A01022515, 2021M3A9E4080780 (National Research Foundation of Korea; to C.-W.W.), 2E30410-20-085 (KIST Institutional Program; to C.-W.W.).

References

Khemakhem, I., Kingma, D., Monti, R., & Hyvarinen, A. (2020). Variational Autoencoders and Nonlinear ICA: A Unifying Framework. In S. Chiappa & R. Calandra (Eds.), *Proceedings of the Twenty Third International Conference on Artificial Intelligence* *and Statistics* (Vol. 108, pp. 2207–2217). PMLR.

Kim, J.-H., Zhang, Y., Han, K., Wen, Z., Choi, M., & Liu, Z. (2021). Representation learning of resting state fMRI with variational autoencoder. *NeuroImage*, *241*, 118423.

Kingma, D. P., & Welling, M. (2013). Auto-Encoding Variational Bayes. In *arXiv [stat.ML]*. arXiv. http://arxiv.org/abs/1312.6114v10

Kingma, & Mohamed. (2014). Semi-supervised learning with deep generative models. *Advances in Neural Information Processing Systems*. https://proceedings.neurips.cc/paper/2014/hash/d523773c6b194f37b938d340d5d02232-Abstract.html

Lim, J., Ryu, S., Kim, J. W., & Kim, W. Y. (2018). Molecular generative model based on conditional variational autoencoder for de novo molecular design. *Journal of Cheminformatics*, *10*(1), 31.

Lucas, J., Tucker, G., Grosse, R. B., & Norouzi, M. (2019). Don't blame the ELBO! A linear VAE perspective on posterior collapse. *Advances in Neural Information Processing Systems*, *32*. https://proceedings.neurips.cc/paper/2019/hash/7e3315fe390974fcf25e44a9445bd821-Abstract.html

Rolinek, M., Zietlow, D., & Martius, G. (2019). Variational autoencoders pursue pca directions (by accident). *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12406–12415.

Zietlow, D., Rolinek, M., & Martius, G. (2021). Demystifying Inductive Biases for (Beta-)VAE Based Architectures. In M. Meila & T. Zhang (Eds.), *Proceedings of the 38th International Conference on Machine Learning* (Vol. 139, pp. 12945–12954). PMLR.

Yarkoni, T., Poldrack, R. A., Nichols, T. E., Van Essen, D. C., & Wager, T. D. (2011). Large-scale automated synthesis of human functional neuroimaging data. *Nature methods*, *8*(8), 665-670.