Disentangling Intermediate Representations in Sound-to-Event DNNs using Invertible Flow Models

Tim Dick (timdi@icloud.com), Enrique Hortal Quesada (enrique.hortal@maastrichtuniversity.nl)
Department of Advanced Computing Sciences, Maastricht University
Paul-Henri Spaaklaan 1, Maastricht, 6229 EN Limburg, The Netherlands

Alexia Briassouli (a.briassouli@utwente.nl)
Faculty of Electrical Engineering, Mathematics and Computer Science, University of Twente
Drienerlolaan 5, 7522 NB Enschede, The Netherlands

Elia Formisano (e.formisano@maastrichtuniversity.nl)
Department of Cognitive Neuroscience, Maastricht University
Oxfordlaan 55, Maastricht, 6229 EV Limburg, The Netherlands

Abstract

Neural representations derived from fMRI responses to natural sounds within non-primary auditory cortical regions mirror those found in the intermediate layers of deep neural networks (DNNs) trained for sound recognition. However, the underlying characteristics of these representations remain elusive. In this study, we investigate the nature of these intermediate representations employing a *disentangling invertible flow model*. We recorded a novel dataset of natural sounds, designed to probe the hypothesis that sound-to-event DNNs encode distinct basic sound generation mechanisms (*human actions*) and source properties (*object materials*) independently within their intermediate layers. To simulate brain responses to these natural sounds, we utilized the layer-by-layer activation of a convolutional DNN (Yamnet), pre-trained to categorize sound spectrograms into semantic categories. Crucially, through systematic manipulations of the obtained latent representations using the disentangling invertible flow model, we demonstrate predictable effects in the DNN's output. This *in silico* demonstration offers a promising avenue for subsequent neuroscientific *in vivo* experimentation. Code available at https://github.com/TimHenry1995/LatentAudio.

Keywords: Machine Learning; Latent Space Disentanglement; Yamnet; Auditory Processing; Normalizing Flow

Introduction

How does the human brain recognize objects and events from sounds? Neuroscience research suggests that non-primary auditory regions in the Superior Temporal Gyrus (STG) play a pivotal role in transforming acoustic representations of natural sounds into semantic representations (Hjortkjær, Kassuba, Madsen, Skov, & Siebner, 2018; Norman-Haignere & McDermott, 2018). Intriguingly, recent studies have revealed that functional magnetic resonance imaging (fMRI) responses to natural sounds in these regions are better explained by convolutional Deep Neural Networks (DNNs) trained to recognize natural sounds than by various other acoustic and semantic models (Kell, Yamins, Shook, Norman-Haignere, & McDermott, 2018; Giordano, Esposito, Valente, & Formisano, 2023). Crucially, latent representations in intermediate layers of the examined DNNs (e.g., VGGish, Yamnet; (Hershey et al., 2017)) most closely resemble those derived from fMRI responses, suggesting that STG regions may involve an intermediate acoustic-semantic representation facilitating sound recognition. However, the specific characteristics of this representation remain elusive, and the extent to which it is necessary for subsequent sound recognition has yet to be determined. While inspecting and manipulating intermediate representations within the human brain is challenging, the latent spaces of artificial neural networks can be disentangled and modified using invertible flow models (Esser, Rombach, & Ommer, 2020; Toledo & Antonelo, 2021; Tomar & Rajagopalan, 2022; Higgins et al., 2016). In this study, this approach is used to explore the nature of said representations using a controlled set of sounds related to human actions and object materials. Additionally, the causal relationship between intermediate and semantic representations is investigated *in silico* by analyzing the impact of systematic manipulations of the latent representations on the DNN output.

Methods

Data Collection

The Material and Action Sound (MaAs) data set introduced here consists of 60,000 1-second long sounds recorded from objects made of wood, metal, glass, stone, cardboard and plastic that interacted by means of tapping, rubbing, destructing and whirling. MaAS distinguishes itself from related data sets (Zhang et al., 2017; Guo, Jiang, & Gao, 2022; Soomro, Zamir, & Shah, 2012; Jung & Chi, 2020; Huh, Chalk, Kazakos, Damen, & Zisserman, 2023) due to its complete factorized design that is needed to disentangle Yamnet's latent space.

Latent Space Exploration and Disentanglement

Materials and actions were decoded from each of Yamnet's 14 layers using 10-fold cross-validated K-Nearest Neighbor (KNN) (Fix & Hodges, 1989; Cover & Hart, 1967) classification. Furthermore, an invertible flow model Esser et al. (2020) was calibrated to disentangle the latent space. This model is

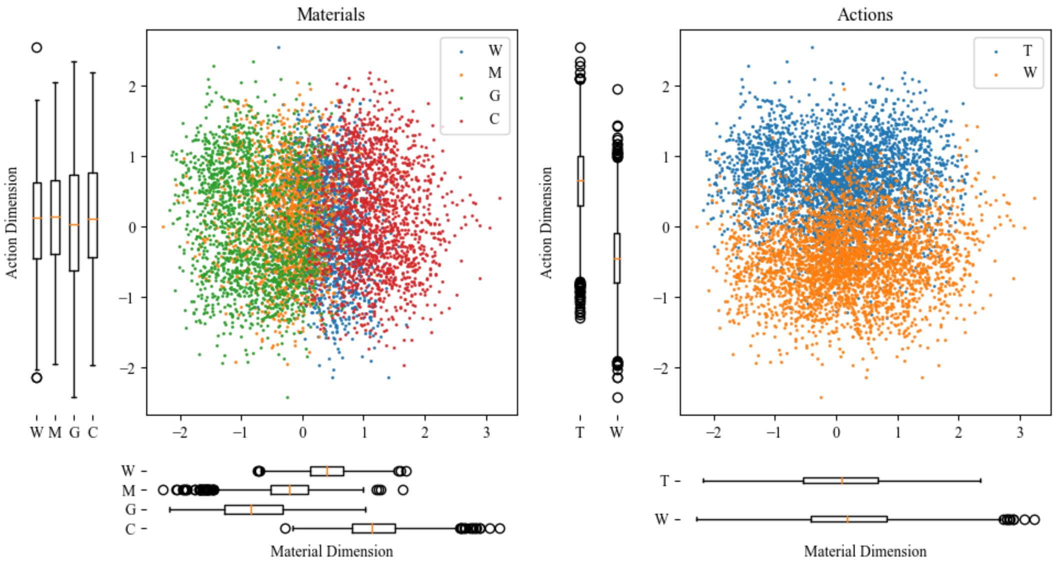

Figure 1: **Latent Space Disentanglement:** Each dot is one MaAS sound, represented by the first two dimensions of the disentangled latent space. Left and right panels show the same dots, except for coloring. Material abbreviations: W (wood), M (metal), G (glass), C (cardboard). Action abbreviations: T (tapping), W (whirling).

a composition of invertible non-linear transformations whose output is supervised to encode material variation only along one dimension, action variation only along a second dimension and all residual variation only along the remaining dimensions. The output is incentivised to be a multivariate normal distribution with clusters along the material dimension for sounds of same material and clusters along the action dimension for sounds of same action.

Latent Space Manipulation

Due to the invertibility of the calibrated flow model, it was possible to map perturbations in the disentangled latent space back onto the original latent space and continue Yamnet's downstream processing. In particular, sounds whose value on the material dimension of the disentangled latent space was atypical (i.e. away from their own material's cluster mean) were made more similar to their own material or another material by perturbing said value. Yamnet's tendency to assign semantic classes typical with either material to the perturbed sound were then measured in Yamnet's final layer. Analogous steps were taken for changes along the action dimension.

Results

Latent Space Exploration and Disentanglement

The decoding of material and action classes by means of KNN was found to be most accurate for intermediate layer 9. This superiority was statistically significant when compared to layers 1 and 14 ($p < 0.01$,t-test, Bonferroni corrected) The disentanglement of materials and actions is shown in Figure 1 for the respective four and two most separable classes.

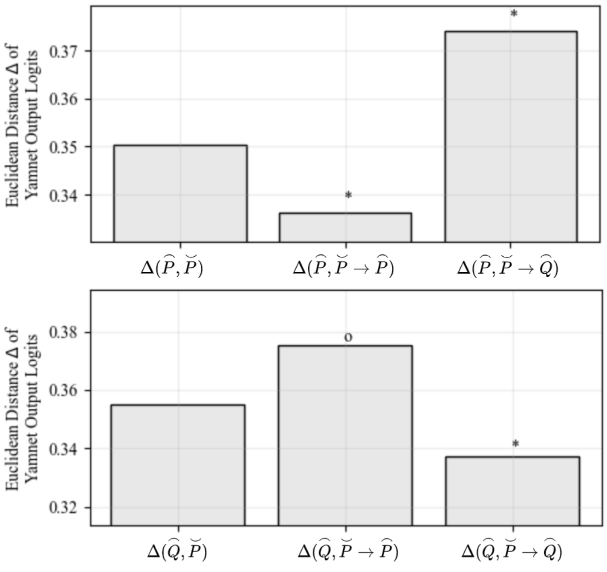

Figure 2: **Latent Transfer:** Effect of transferring an atypical ($\smile$) sound to a typical ($\frown$) one of the same (P) or different (Q) material. $*$ = significant at Bonferroni corrected $\alpha = 5\%$.

Latent Space Manipulation

As shown in Figure 2 (top), making an atypical sound $\overset{\smile}{P}$ more similar to a typical sound $\overset{\frown}{P}$ of same material in the latent space reduces their distance in Yamnet's semantic space (middle vs. left bar). Yet, making $\overset{\smile}{P}$ more similar to a typical sound of another material $\overset{\frown}{Q}$ makes it diverge from $\overset{\frown}{P}$ in the semantic space (right vs. left bar). Importantly, the effect is reversed when the distances are measured with respect to $\overset{\frown}{Q}$ (bottom half of Figure 2) such that moving $\overset{\smile}{P}$ to $\overset{\frown}{Q}$ in the latent space makes the atypical $\overset{\smile}{P}$ sound more typical for the semantic classes of $\overset{\frown}{Q}$ (right vs. left bar). The same observations hold for action perturbations (not shown here). These comparisons are statistically significant at $\alpha = 0.05$.

Conclusions

Our analyses indicate that semantic supervision fosters the emergence of independent material and action representations in sound-to-event DNNs, thereby mirroring empirical (fMRI) findings in human auditory cortex Hjortkjær et al. (2018). Finally, our *in silico* manipulation of the latent space suggests a causal link between these representations and higher level semantics *in silico* and may constitute an important step to understanding sound event recognition in *in vivo*.

Acknowledgments

This work was funded by the Dutch Research Council (NWO 406.20.GO.030 to EF).

References

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE transactions on information theory*, *13*(1), 21–27.

Esser, P., Rombach, R., & Ommer, B. (2020). A disentangling invertible interpretation network for explaining latent representations. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 9223–9232).

Fix, E., & Hodges, J. L. (1989). Discriminatory analysis. nonparametric discrimination: Consistency properties. *International Statistical Review/Revue Internationale de Statistique*, *57*(3), 238–247.

Giordano, B. L., Esposito, M., Valente, G., & Formisano, E. (2023). Intermediate acoustic-to-semantic representations link behavioral and neural responses to natural sounds. *Nature Neuroscience*, 1–9.

Guo, H., Jiang, H., & Gao, M. (2022). Research on sound source material recognition technology in indoor geotechnical inspection. *Geofluids*, *2022*.

Hershey, S., Chaudhuri, S., Ellis, D. P., Gemmeke, J. F., Jansen, A., Moore, R. C., ... others (2017). Cnn architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 131–135).

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., ... Lerchner, A. (2016). beta-vae: Learning basic visual concepts with a constrained variational framework. In *International conference on learning representations.*

Hjortkjær, J., Kassuba, T., Madsen, K. H., Skov, M., & Siebner, H. R. (2018). Task-modulated cortical representations of natural sound source categories. *Cerebral Cortex*, *28*(1), 295–306.

Huh, J., Chalk, J., Kazakos, E., Damen, D., & Zisserman, A. (2023). Epic-sounds: A large-scale dataset of actions that sound. In *Icassp 2023-2023 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 1–5).

Jung, M., & Chi, S. (2020). Human activity classification based on sound recognition and residual convolutional neural network. *Automation in Construction*, *114*, 103177.

Kell, A. J., Yamins, D. L., Shook, E. N., Norman-Haignere, S. V., & McDermott, J. H. (2018). A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. *Neuron*, *98*(3), 630–644.

Norman-Haignere, S. V., & McDermott, J. H. (2018). Neural responses to natural and model-matched stimuli reveal distinct computations in primary and nonprimary auditory cortex. *PLoS Biology*, *16*, e2005127.

Soomro, K., Zamir, A. R., & Shah, M. (2012). Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*.

Toledo, R. S., & Antonelo, E. A. (2021). Face reconstruction with variational autoencoder and face masks. *arXiv preprint arXiv:2112.02139*.

Tomar, S. S., & Rajagopalan, A. (2022). Latents2segments: Disentangling the latent space of generative models for semantic segmentation of face images. *arXiv preprint arXiv:2207.01871*.

Zhang, Z., Li, Q., Huang, Z., Wu, J., Tenenbaum, J., & Freeman, B. (2017). Shape and material from sound. *Advances in Neural Information Processing Systems*, *30*.