Mapping Modular Processing of Compressed Videos Across Human Visual Cortex

Christina Sartzetaki (c.sartzetaki@uva.nl), Iris I.A. Groen (i.i.a.groen@uva.nl)
Informatics Institute, University of Amsterdam, The Netherlands

Abstract

How the brain makes sense of the constant stream of visual information it receives largely remains a mystery. One prominent idea is that it evolved specialized pathways for sparse cortical engagement, and those can be accurately captured with handcrafted features; however, deep neural network (DNN) features overall align better with brain representations. In this work we study the brain alignment of a multi-pathway DNN that leverages compressed video formats, and partition the variance captured between its three modular components across visual brain regions recorded with fMRI during video stimuli. We find that its components map well to known brain pathways, and that it captures overall more variance than a 3D convolutional network. Achieved using only existing features in the compressed format, this points to the ineffectiveness of conventional full-frame processing for explaining brain responses to dynamic stimuli and to compression as a potential solution.

Keywords: variance partitioning; dynamic natural stimuli; fMRI; action recognition DNNs

Introduction

In recent years, traditional computational models of human vision using theory-driven hand-crafted features are rivaled by deep neural networks (DNNs) with learned task-optimized features. Hand-crafted video features distinctly map to different pathways in visual cortex during viewing of natural dynamic stimuli (Bartels, Zeki, & Logothetis, 2008; Nishimoto & Gallant, 2011), but currently DNN features achieve the highest alignment to brain responses in static image stimuli benchmarks (Schrimpf et al., 2018; Kriegeskorte, 2015) and for dynamic stimuli they also align differently depending on their temporal modeling (Sartzetaki, Roig, Snoek, & Groen, 2025). However, DNNs trained on top of hand-crafted features such as optical flow still explain unique variance compared to larger end-to-end networks (Karimi & Anzellotti, 2024).

Accurate optical flow computation for every pixel or separate processing of all RGB frames in a video are implausible in both biological and artificial intelligence, as for any limited-capacity system efficient data compression is needed, and feasible due to the rich statistical structure of most signals (Bates & Jacobs, 2020). Eliminating redundancy in sensory information results in sparse coding engaging only a small portion of cortex simultaneously (Olshausen & Field, 2004), and for prolonged static stimuli we also observe response reductions (Zhou, Benson, Kay, & Winawer, 2018; Groen et al., 2022; Brands et al., 2024) reflecting compressive temporal summation. In computer vision, compressed action recognition models (Wu et al., 2018; Chen & Ho, 2022) take advantage of the commonly used storing and transmission format, MPEG4 compression. They dedicate less resources to processing full RGB I-frames, and also compute representations on the intermediate P-frames, comprised of motion vectors and residuals (proxies for optical flow and edge features).

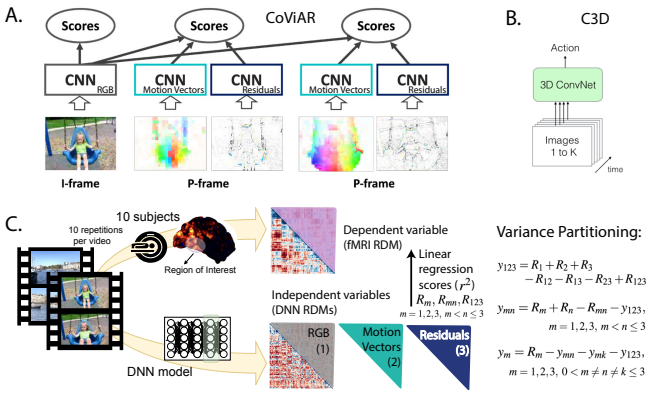

Figure 1: A. Information flow in the compressed action recognition model, CoViAR (Wu et al., 2018), and B. a 3D CNN. C. Pipeline for Variance Partitioning between CoViAR's three modular CNNs on video fMRI data (Lahner et al., 2024).

Motivated by this, here we compute the representational alignment of one such model, CoViAR (Wu et al., 2018), to the human brain watching videos, and compare to 3D convolutional models as baseline. We find that it explains more unique variance than a 3DCNN and that its modular components neatly map to the respective functional brain pathways.

Methods

Neural dataset We employ the Bold Moments Dataset (BMD) (Lahner et al., 2024) consisting of whole-brain 3T fMRI recordings from 10 subjects watching 1102 3s videos from the Moments in Time (Monfort et al., 2019) video dataset. For each subject, 1000 videos were shown for 3 repetitions whereas 102 videos were shown for 10 repetitions - here we use this latter subset (whose videos are sensibly representative of the whole set). We use preprocessed data provided by Lahner et al. (2024), concatenating voxels for each available brain ROI across hemispheres, as well as dorsal and ventral V1 and V2.

Model training and feature extraction We train the models of the three input streams (shown in **Fig. 1**[1]), RGB (Resnet-152), Motion vectors (Resnet-18), and Residuals (Resnet-18), for action recognition on the dataset UCF101 (Soomro, Zamir, & Shah, 2012) using the codebase and hyperparameters provided in Wu et al. (2018), and successfully reproduce their results. The C3D and TSM R50 MobileOne

[1] Figure parts A and B adapted from (Wu et al., 2018) and (Simonyan & Zisserman, 2014) respectively.

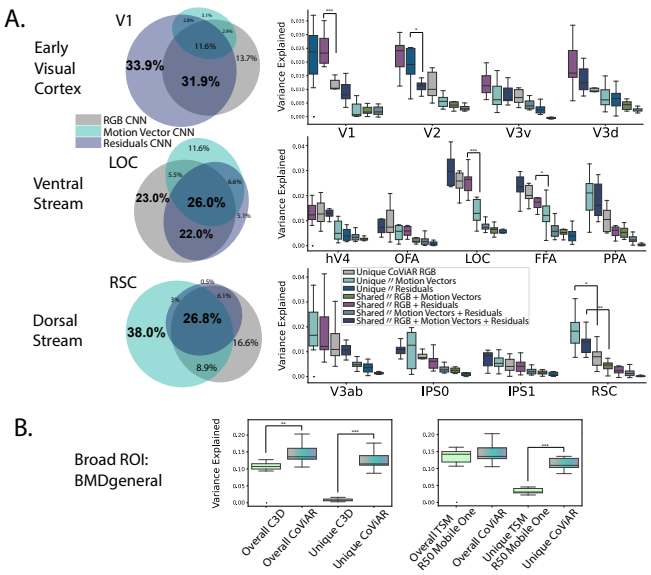

Figure 2: A. Variance Partitioning (VP) between the CoViAR components in different ROIs displayed as Euler diagrams (left) and r^2 score distributions across participants (right). B. VP between CoViAR and C3D, as well as a more high-aligned 3D CNN. Stars indicate pairwise significant differences.

models are ported from the mmaction2[2] library, trained on UCF101 and Kinetics-400 respectively. Each model expects a clip of specific length, so we average the features across all sub-clips of a video. We extract features from all higher-level blocks in the models, also including the final fully connected classification layer, and flatten the features after extraction to produce a single one-dimensional feature vector per layer.

Variance Partitioning on RDMs We construct Representational Dissimilarity Matrices (RDMs) (Kriegeskorte, Mur, & Bandettini, 2008) with Pearson correlation from the voxel vectors of each ROI and subject (after averaging across repetitions), and from the feature vectors of each model layer (after reducing the dimensionality to 100PCs with Principal Component Analysis). We choose the model layer RDM that achieves the highest average Spearman correlation with the subject RDMs to perform Variance Partitioning (VP) (Legendre, 2008). For VP between three models we fit a linear regression from each model RDM m to each subject RDM and compute the r^2 score R_m, and a linear regression from each combination of model RDMs m, n to each subject RDM and compute the r^2 scores R_{mn} and R_{123}. We then partition the variance to obtain shared and unique variance partitions as shown in **Fig. 1.C.** We utilize the Net2Brain python library (Bersch et al., 2025) for parts of the pipeline, and the eulerAPE program[3] for the diagrams of **Fig. 2**. We compute statistical significance of variance distributions against zero with a one-sample t-test, and pairwise significance between two variance distributions with Welch's t-test.

[2] https://mmaction2.readthedocs.io

[3] https://www.eulerdiagrams.com/eulerAPE/v2/

Results

In **Fig. 2.A(left)** we show results for Variance Partitioning (VP) between the RGB, Motion Vector, and Residuals CNN streams of the CoViAR model in the form of Euler (elliptical Venn) diagrams, for selected brain ROIs from the Early Visual Cortex (EVC), Ventral stream, and Dorsal stream. Partitions are shown in percentages of the total variance explained by all the models together (the union of the ellipses). In the EVC, the Residuals CNN uniquely accounts for the largest amount of variance in V1 (33.9%), followed by the variance shared between the Residuals and the RGB CNN (31.9%). In the Ventral stream, similar amounts of variance are explained by the shared contributions of all three CNNs (26%), the unique contributions of the RGB CNN (23%), and the shared contributions of the Residuals and the RGB CNNs (22%). In the Dorsal stream, the Motion Vector CNN takes a clear lead in uniquely explaining the most variance (38%), followed by the variance shared between all three CNNs (26.8%). These results are shown in more detail in **Fig. 2.A(right)**, including the distributions of r^2 scores across subjects and in all ROIs, as well as selected pairwise significances. All score distributions are significant against zero. The total variance accounted for by all partitions together amounts to $1/3 - 1/2$ of the lower noise ceiling present in the data, depending on the ROI.

In **Fig. 2.B** we show results for two 4-way VP analyses, between the three CoViAR models and two 3DCNNs, in the broad ROI BMDgeneral (Lahner et al., 2024). We first observe that the overall variance explained by all three CoViAR CNNs (R_{123}) is significantly higher (p=0.009) than that explained by the C3D model (R_4). We further investigate by subtracting any shared variance between the overall CoViAR variance and C3D, i.e. $R_{123} - y_{14} - y_{24} - y_{34} + y_{124} + y_{134} + y_{234} - y_{1234}$, and comparing to the unique variance of C3D y_4, and there find an even larger advantage of CoViAR. Interestingly, we find the same when repeating the VP for TSM R50 MobileOne, shown as highly brain-aligned in Sartzetaki et al. (2025).

Discussion

Leveraging the already available compressed video format to account for differences between consecutive frames, CoViAR's CNN streams for RGB, Motion Vectors, and Residuals map well to human brain regions known for processing these respective types of information; the EVC's biggest variance partition Residuals resemble edge features, the RGB and Residuals can contribute to object recognition in the Ventral stream, while Motion Vectors can account for dynamic features found in Dorsal areas. We note a lack of high cortex engagement for uniquely computing RBG features, while the RGB-only 3DCNN - though computing temporal features - also fails to capture large unique variance compared to CoViAR. These results may collectively point to the ineffectiveness of RGB features during dynamic tasks like video-watching. At the same time, they are the most expensive to compute of the three, as is 3D convolution compared to 2D, so a multi-stream solution might naturally emerge for sparsity and efficiency reasons (Olshausen & Field, 2004).

Acknowledgments

This work is supported by an ELLIS Amsterdam Unit grant to IIAG.

References

Bartels, A., Zeki, S., & Logothetis, N. K. (2008). Natural vision reveals regional specialization to local motion and to contrast-invariant, global flow in the human brain. *Cerebral cortex*, *18*(3), 705–717.

Bates, C. J., & Jacobs, R. A. (2020). Efficient data compression in perception and perceptual memory. *Psychological review*, *127*(5), 891.

Bersch, D., Vilas, M. G., Saba-Sadiya, S., Schaumlöffel, T., Dwivedi, K., Sartzetaki, C., ... Roig, G. (2025). Net2brain: A toolbox to compare artificial vision models with human brain responses. *Frontiers in Neuroinformatics*, *19*, 1515873.

Brands, A. M., Devore, S., Devinsky, O., Doyle, W., Flinker, A., Friedman, D., ... Groen, I. I. A. (2024). Temporal dynamics of short-term neural adaptation across human visual cortex. *PLOS Computational Biology*, *20*(5), e1012161.

Chen, J., & Ho, C. M. (2022). Mm-vit: Multi-modal video transformer for compressed video action recognition. In *Proceedings of the ieee/cvf winter conference on applications of computer vision* (pp. 1910–1921).

Groen, I. I., Piantoni, G., Montenegro, S., Flinker, A., Devore, S., Devinsky, O., ... others (2022). Temporal dynamics of neural responses in human visual cortex. *Journal of Neuroscience*, *42*(40), 7562–7580.

Karimi, H., & Anzellotti, S. (2024). Comparing representations in static and dynamic vision models to the human brain. In *Unireps: 2nd edition of the workshop on unifying representations in neural models.*

Kriegeskorte, N. (2015). Deep neural networks: a new framework for modeling biological vision and brain information processing. *Annual review of vision science*, *1*(1), 417–446.

Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008). Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, *2*, 249.

Lahner, B., Dwivedi, K., Iamshchinina, P., Graumann, M., Lascelles, A., Roig, G., ... others (2024). Modeling short visual events through the bold moments video fmri dataset and metadata. *Nature communications*, *15*(1), 6241.

Legendre, P. (2008). Studying beta diversity: ecological variation partitioning by multiple regression and canonical analysis. *Journal of plant ecology*, *1*(1), 3–8.

Monfort, M., Andonian, A., Zhou, B., Ramakrishnan, K., Bargal, S. A., Yan, T., ... others (2019). Moments in time dataset: one million videos for event understanding. *IEEE transactions on pattern analysis and machine intelligence*, *42*(2), 502–508.

Nishimoto, S., & Gallant, J. L. (2011). A three-dimensional spatiotemporal receptive field model explains responses of area mt neurons to naturalistic movies. *Journal of Neuroscience*, *31*(41), 14551–14564.

Olshausen, B. A., & Field, D. J. (2004). Sparse coding of sensory inputs. *Current opinion in neurobiology*, *14*(4), 481–487.

Sartzetaki, C., Roig, G., Snoek, C. G. M., & Groen, I. (2025). One hundred neural networks and brains watching videos: Lessons from alignment. In *The thirteenth international conference on learning representations.* Retrieved from https://openreview.net/forum?id=LM4PYXBId5

Schrimpf, M., Kubilius, J., Hong, H., Majaj, N. J., Rajalingham, R., Issa, E. B., ... others (2018). Brain-score: Which artificial neural network for object recognition is most brain-like? *BioRxiv*, 407007.

Simonyan, K., & Zisserman, A. (2014). Two-stream convolutional networks for action recognition in videos. *Advances in neural information processing systems*, *27*.

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*.

Wu, C.-Y., Zaheer, M., Hu, H., Manmatha, R., Smola, A. J., & Krähenbühl, P. (2018). Compressed video action recognition. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 6026–6035).

Zhou, J., Benson, N. C., Kay, K. N., & Winawer, J. (2018). Compressive temporal summation in human visual cortex. *Journal of Neuroscience*, *38*(3), 691–709.