Straightening of Natural Visual Sequences in Video DNNs: the Role of Locality and Temporal Coherence

Anne W. Zonneveld (a.w.zonneveld@uva.nl)
Institute of Informatics, University of Amsterdam, the Netherlands

Pascal Mettes * (p.s.m.mettes@uva.nl)
Institute of Informatics, University of Amsterdam, the Netherlands

Iris I. A. Groen * (i.i.a.groen@uva.nl)
Institute of Informatics, University of Amsterdam, the Netherlands

* Shared senior author

Abstract

Predictions about future states of the world play an important role in guiding human behavior. In vision, internal representations are straightened relative to input space, supporting linear extrapolation and predictability. While DNNs are promising models of human visual processing, standard image-trained architectures lack this property. We assessed straightening, quantified as a reduction in curvature in feature vs. pixel space, across 19 different video DNNs, including both CNNs and Transformers. Straightening occurred in late CNN layers, but was absent in Transformers. Critically, models with global attention lacked temporal coherence in feature space, a prerequisite for straightening. These findings suggest that temporally localized processing, like 3D convolutions, enables brain-like invariant representations, whereas Transformers, despite their strong performance, may rely on mechanisms that diverge from biological vision.

Keywords: temporal straightening; deep neural network; video

Introduction

In human vision, retinal input is transformed along the cortical visual hierarchy into increasingly invariant representations, enabling manifold untangling for classification (DiCarlo & Cox, 2007). This can be extended to the temporal domain: visual changes over time form temporal trajectories that become untangled, or *straightened,* in human perception (Hénaff et al., 2019) and macaque V1 (Hénaff et al., 2021) compared to pixel space (**Fig. 1A**). While deep neural networks (DNNs) are promising models of human vision (Conwell et al., 2024; Schrimpf et al., 2020), standard image-trained DNNs do not show straightening (Harrington et al., 2022; Hénaff et al., 2019; Toosi & Issa, 2023). Straightened representations may be beneficial for computer vision by enhancing predictability through linear extrapolation, supporting future state prediction (Niu et al., 2024; Toosi & Issa, 2023), robustness (Harrington et al., 2022; Niu et al., 2024; Toosi & Issa, 2023) and classification confidence (Harrington et al., 2022). These prior works mainly focused on image-trained DNNs, treating frames independently and therefore not accounting for the potential of history-dependent mechanisms afforded by temporal context unique to video. In contrast, video DNNs, either with or without explicit temporal modeling (e.g. 3D convolutions or spatiotemporal attention), may exploit the regularities of natural videos, enabling straightening. Here, we demonstrate that straightening naturally emerges in late layers of CNN-based video DNNs.

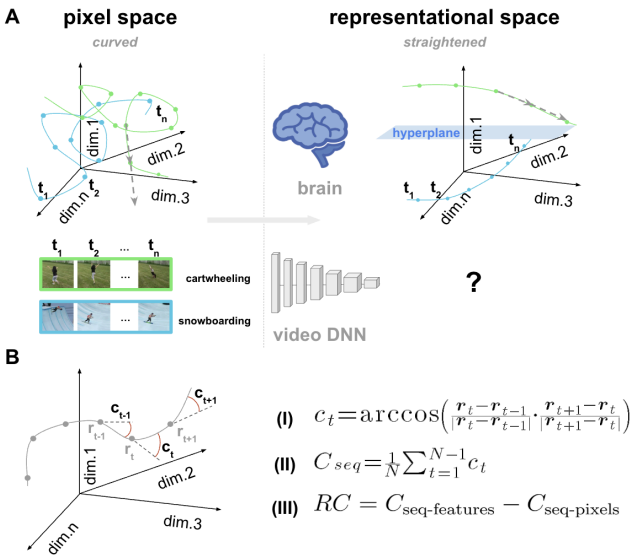

Fig. 1. A. Temporal straightening hypothesis **B.** Curvature calculation as defined by Hénaff et al. (2019)

Furthermore, we find that global attention disrupts temporal coherence, which is necessary for straightening. This suggests that temporally localized processing, e.g. via 3D convolutions, supports brain-like invariant representations, whereas Transformers may employ mechanisms that differ from biological vision.

Methods

Straightening metric

Straightening is estimated as the reduction of curvature of the temporal trajectory of representations (e.g., brain activations or DNN features) compared to the input space, in our case video frames (**Fig. 1A**). Local curvature c_t (**Fig. 1B**; I) is defined as the angle between two vectors representing the difference between the representations *r* of consecutive points in time (Hénaff et al., 2019), averaged over the sequence to obtain global curvature (**Fig. 1B**; II). Lower curvature indicates a straighter trajectory. 'Straightening' refers to negative *relative curvature* (RC), i.e., global curvature in DNN feature space compared to pixel space (**Fig. 1B**; III).

Model choice

We evaluated 19 video DNNs: 13 CNNs and 6 Transformers, all trained on Kinetics400 (Kay et al., 2017) for action classification. We also included one image model (ResNet-50 (He et al., 2016)), trained on ImageNet (Deng et al., 2009) for object classification. Models were retrieved from the PyTorchVideo (Fan et al., 2021) and mmaction2 library (MMAction2 Contributors, 2020).

Model feature extraction

We followed preprocessing guidelines from the respective model sources. Each video model uses a specific sampling strategy, defined by the number of input frames and sampling rate (F x R). The image model followed a common strategy of 8x8. Features were extracted for early, mid and late activation layers, with dimensions of [channels x time x height x width].

Evaluation

We evaluated video model straightening using 1102 three-second videos from the Bold Moments Dataset (Lahner et al., 2024). Curvature was computed for each video across time in both (model-specific) pixel space and in feature space, for all layers and models of interest. We conducted two statistical tests. First, to establish evidence of temporal straightening, a one-sided one-sample t-test comparing RC to zero ($p < 0.05$, FDR-corrected). Second, to investigate potential underlying causes of (the absence) of straightening, we tested for temporal coherence, a prerequisite for straightening, by shuffling features along the time axis (1000 iterations). If model features are temporally coherent, curvature for unshuffled features should be lower than that of shuffled features, as shuffling disrupts temporal continuity and leads to irregular trajectories in feature space. Temporal coherence was inferred if original curvature was significantly lower than the permuted distribution (one-sided test, $p < 0.05$, FDR-corrected).

Results

Straightening occurs in late layers of CNNs

In early layers, only one model (SlowFast; Feichtenhofer et al., 2019) showed significant straightening; no models showed straightening in middle layers (**Fig. 2A**). Several CNNs exhibited straightening in late layers (Slow (only) (Feichtenhofer et al., 2019), X3D S/M/L (Feichtenhofer, 2020), TANet (Liu et al., 2021), I3D (Carreira & Zisserman, 2017), R2+1D (Tran et al., 2018), Omnisource (Duan et al., 2020), while no Transformers did. CNNs typically exhibited a positive RC from early to middle layers, followed by a decrease, resulting in negative RC (i.e. straightening), in late layers. This suggests an initial increased responsiveness to fine-grained temporal shifts before information is consolidated into temporally invariant representations in later layers. In contrast, most Transformers maintained a positive RC across layers (**Fig. 2B**). Notably, video CNNs that perform only temporal feature aggregation without genuine temporal modeling, i.e. 'image-based' video models (coined by Sartzetaki et al. (2024)), such as C2D (X. Wang et al., 2018) and TSN (L. Wang et al., 2016), did not exhibit straightening.

Global attention disrupts temporal coherence

Temporal coherence was consistently observed across all layers of CNNs, but not across all layers in Transformers (**Fig. 2A**). Coherence was preserved in local-attention models (MViT; Fan et al., 2021), Swin

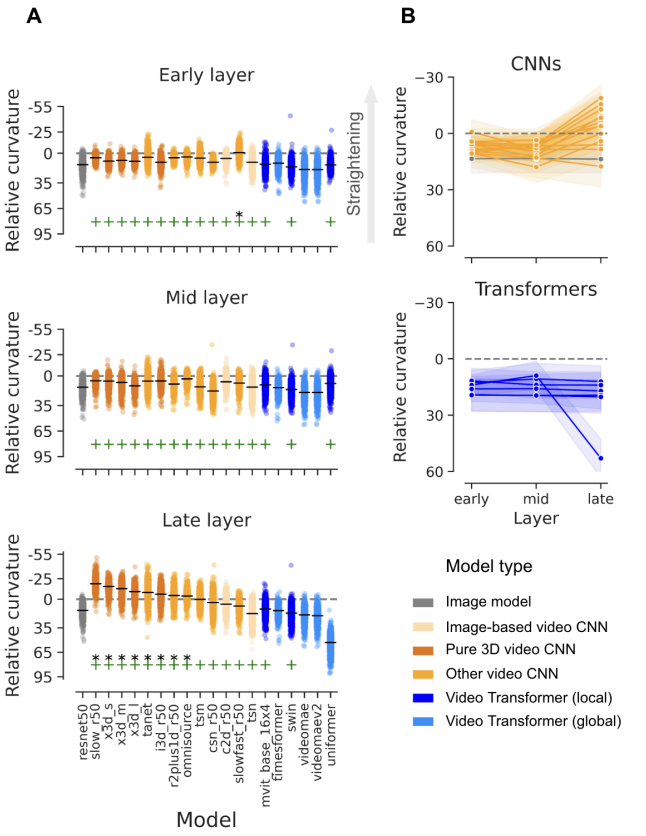

Fig. 2. A. RC in different layers for varying video DNNs. * = one-sided t-test against 0, + = temporal coherence test. **B.** Mean RC over layers in CNN-based vs transformer based video DNNs.

(Liu et al., 2022), and early/mid layers of UniFormer (Li et al., 2022), but absent in global-attention models (TimeSformer; Bertasius et al., 2021) and late layers of UniFormer (Li et al., 2022). This suggests that global attention interferes with the temporal alignment of the representational and input space by enabling non-localized spatiotemporal relationships across varying timescales, which may explain why global attention Transformers fail to exhibit straightening.

Conclusion

First, we found that temporal straightening only occurs in late layers of CNNs. Second, we demonstrated that global attention disrupts temporal coherence. Taken together, these findings suggest that temporal coherence is a necessary, but not sufficient, condition for representational straightening in video DNNs. Specifically, our results highlight a critical role of *locality* of temporal integration to enable straightening: local operations, such as 3D convolutions, support the emergence of invariant representations over time, potentially enhancing robustness (Harrington et al., 2022; Niu et al., 2024; Toosi & Issa, 2023) and classification confidence (Harrington et al., 2022). While Transformers outperform CNNs in accuracy and align more closely with human error patterns (Tuli et al., 2021), our findings suggest that their success may rely on mechanisms that diverge from principles of biological vision.

Acknowledgements

This work was supported by the UvA Data Science Centre, as part of the Human Aligned Video AI Lab. Furthermore I would like to thank Pablo Oyarzo for insightful discussions about the project.

References

Bertasius, G., Wang, H., & Torresani, L. (2021). Is space-time attention all you need for video understanding? *arXiv:2102.05095*. https://doi.org/10.48550/arXiv.2102.0509

Carreira, J., & Zisserman, A. (2017). Quo vadis, action recognition? A new model and the Kinetics dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 6299–6308). https://openaccess.thecvf.com/content_cvpr_2017/html/Carreira_Quo_Vadis_Action_CVPR_2017_paper.html

Conwell, C., Prince, J. S., Kay, K. N., Alvarez, G. A., & Konkle, T. (2024). A large-scale examination of inductive biases shaping high-level visual representation in brains and machines. *Nature Communications, 15*(1), 9383. https://doi.org/10.1038/s41467-024-53147-y

Deng, J., Dong, W., Socher, R., et al. (2009). ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 248–255). https://doi.org/10.1109/CVPR.2009.5206848

DiCarlo, J. J., & Cox, D. D. (2007). Untangling invariant object recognition. *Trends in Cognitive Sciences, 11*(8), 333–341. https://doi.org/10.1016/j.tics.2007.06.010

Duan, H., Zhao, Y., Xiong, Y., Liu, W., & Lin, D. (2020). Omni-sourced webly-supervised learning for video recognition. *arXiv:2003.13042*. https://doi.org/10.48550/arXiv.2003.13042

Fan, H., Murrell, T., Wang, H., Alwala, K. V., Li, Y., Li, Y., ... & Feichtenhofer, C. (2021, October). PyTorchVideo: A deep learning library for video understanding. In *Proceedings of the 29th ACM international conference on multimedia* (pp. 3783-3786). https://doi.org/10.1145/3474085.3478329

Fan, H., Xiong, B., Mangalam, K., et al. (2021). Multiscale vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 6824–6835). https://openaccess.thecvf.com/content/ICCV2021/html/Fan_Multiscale_Vision_Transformers_ICCV_2021_paper.html

Feichtenhofer, C. (2020). X3D: Expanding architectures for efficient video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 203–213). https://openaccess.thecvf.com/content_CVPR_2020/html/Feichtenhofer_X3D_Expanding_Architectures_for_Efficient_Video_Recognition_CVPR_2020_paper.html

Feichtenhofer, C., Fan, H., Malik, J., & He, K. (2019). SlowFast networks for video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 6202–6211). https://openaccess.thecvf.com/content_ICCV_2019/html/Feichtenhofer_SlowFast_Networks_for_Video_Recognition_ICCV_2019_paper.html

Harrington, A., DuTell, V., Tewari, A., et al. (2022). Exploring perceptual straightness in learned visual representations. In *Proceedings of the International Conference on Learning Representations*. https://openreview.net/forum?id=4cOfD2qL6T

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 770–778). https://openaccess.thecvf.com/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html

Hénaff, O. J., Bai, Y., Charlton, J. A., et al. (2021). Primary visual cortex straightens natural video trajectories. *Nature Communications, 12*, 5982. https://doi.org/10.1038/s41467-021-25939-z

Hénaff, O. J., Goris, R. L. T., & Simoncelli, E. P. (2019). Perceptual straightening of natural videos. *Nature Neuroscience, 22*(6), 984–991. https://doi.org/10.1038/s41593-019-0377-4

Kay, W., Carreira, J., Simonyan, K., et al. (2017). The Kinetics human action video dataset. *arXiv:1705.06950*. https://doi.org/10.48550/arXiv.1705.06950

Lahner, B., Dwivedi, K., Iamshchinina, P., et al. (2024). BOLD Moments: Modeling short visual events through a video fMRI dataset and metadata. *bioRxiv*. https://doi.org/10.1101/2023.03.12.530887

Li, K., Wang, Y., Gao, P., et al. (2022). UniFormer: Unified transformer for efficient spatiotemporal representation learning. *arXiv:2201.04676*. https://doi.org/10.48550/arXiv.2201.04676

Liu, Z., Ning, J., Cao, Y., et al. (2022). Video Swin Transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 3202–3211). https://openaccess.thecvf.com/content/CVPR2022/html/Liu_Video_Swin_Transformer_CVPR_2022_paper.html

Liu, Z., Wang, L., Wu, W., et al. (2021). TAM: Temporal adaptive module for video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 13708–13718). https://openaccess.thecvf.com/content/ICCV2021/html/Liu_TAM_Temporal_Adaptive_Module_for_Video_Recognition_ICCV_2021_paper.html

MMAction2 Contributors. (2020). *OpenMMLab's next*

generation video understanding toolbox and benchmark. GitHub. https://github.com/open-mmlab/mmaction2

Monfort, M., Andonian, A., Zhou, B., et al. (2020). Moments in Time Dataset: One million videos for event understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 42*(2), 502–508. https://doi.org/10.1109/TPAMI.2019.2901464

Monfort, M., Pan, B., Ramakrishnan, K., et al. (2022). Multi-Moments in Time: Learning and interpreting models for multi-action video understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 44*(12), 9434–9445. https://doi.org/10.1109/TPAMI.2021.3126682

Niu, X., Savin, C., & Simoncelli, E. P. (2024). Learning predictable and robust neural representations by straightening image sequences. *arXiv:2411.01777*. https://doi.org/10.48550/arXiv.2411.01777

Sartzetaki, C., Roig, G., Snoek, C. G. M., & Groen, I.I.A. (2025). One hundred neural networks and brains watching videos: Lessons from alignment. *Proceedings of the Thirteenth International Conference on Learning Representations*. https://openreview.net/forum?id=LM4PYXBId5

Schrimpf, M., Kubilius, J., Hong, H., et al. (2020). Brain-Score: Which artificial neural network for object recognition is most brain-like? *bioRxiv*. https://doi.org/10.1101/407007

Toosi, T., & Issa, E. B. (2023). Brain-like representational straightening of natural movies in robust feedforward neural networks. *arXiv:2308.13870*. https://doi.org/10.48550/arXiv.2308.13870

Tran, D., Wang, H., Torresani, L., et al. (2018). A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 6450–6459). https://openaccess.thecvf.com/content_cvpr_2018/html/Tran_A_Closer_Look_CVPR_2018_paper.html

Tuli, S., Dasgupta, I., Grant, E., & Griffiths, T. L. (2021). Are convolutional neural networks or transformers more like human vision?. *arXiv:2105.07197*. https://doi.org/10.48550/arXiv.2105.07197

Wang, L., Xiong, Y., Wang, Z., et al. (2016). Temporal segment networks: Towards good practices for deep action recognition. In *Proceedings of the European Conference on Computer Vision* (pp. 20–36). https://doi.org/10.1007/978-3-319-46484-8_2

Wang, X., Girshick, R., Gupta, A., & He, K. (2018). Non-local neural networks. *arXiv:1711.07971*. https://doi.org/10.48550/arXiv.1711.07971