JOINT INTERPRETATION OF REPRESENTATIONS IN NEURAL NETWORKS AND THE BRAIN

Aria Yuan Wang*, Ruogu Lin*, Michael J. Tarr & Leila Wehbe
Carnegie Mellon University
{ariawang, lrg14, michaeltarr, lwehbe}@cmu.edu

ABSTRACT

Recent successes in using representations from deep neural networks to predict brain responses promise to advance our understanding of hierarchical information processing in the primate brain. The productivity of this approach points to a convergence in representation between the brain and artificial neural networks. Given that both systems learn to achieve high levels of performance for real-world vision tasks, we address two questions: i) How far does this convergence extend? ii) What are the factors that influence this convergence? Here we investigate how different choices of tasks and networks can affect the mapping from neural network representations to brain responses. We build stacked voxelwise encoding models and compare prediction performance and stacking weights. Our results demonstrate that these choices may affect correspondences between neural networks and brains, giving rise to varying interpretations of neural responses. Importantly, our results also demonstrate that leveraging our extensive existing knowledge of the brain makes it possible to gain insight into learned representations in artificial neural networks.

1 INTRODUCTION

Advances in neural networks have spurred dramatic improvement in artificial vision systems. While the performance of such systems across a wide variety of vision tasks is impressive, to understand how these networks evolve to achieve high task accuracy remains challenging. Regardless, neural networks, or more specifically, their learned representations, have been useful as proxy models for hierarchical processing in the brain (Agrawal et al., 2014; Yamins et al., 2014; Güçlü & van Gerven, 2015; Kell et al., 2018). Given similarity in the task end goals of both artificial and biological systems, it is not surprising that high-performing systems in both domains share representations despite drastically different physical implementations(Yamins & DiCarlo, 2016). More broadly, we see a similar convergence in many domains, including vision (Agrawal et al., 2014; Yamins et al., 2014; Güçlü & van Gerven, 2015; Schrimpf et al., 2018), audition (Kell et al., 2018), language (Wehbe et al., 2014; Jain & Huth, 2018; Caucheteux & King, 2020; Jain et al., 2020), and both feedforward and recurrent networks (Wehbe et al., 2014; Nayebi et al., 2021).

The "explanatory arrow" has almost always been unidirectional – what can artificial neural networks and their learned representations tell us about brain representations. Implicit in this directionality is the assumption that neural networks are *good* models for neural systems; that is, that the computations implemented in neural networks help us to better understand the "black box" computations realized in different parts of the brain. Here we take a deeper look into how various choices of network in terms of layers and task the network is trained on could affects how well these representation can predict the the brain. Our results indicate that the converse is also possible: facts about the brain can help us to better understand computations and representation in artificial neural networks.

Interest in "interpretable AI" and different methods for visualizing representation in artificial neural networks has exploded over the past several years (Mordvintsev et al., 2015; Olah et al., 2017; 2018; Bau et al., 2017). Yet there are limitations on how much one can learn from visualization of network features - not the least of which is the human tendency to assign a greater semantic meaning and functional relevance to visualizations than otherwise might be warranted. On the other hand, there is a century long history of visual neuroscience on which we can build (Gross, 1994). For example,

Hubel and Wiesel's (Hubel & Wiesel, 1959) elucidation of the response properties of localized receptive fields – a concept that forms the basis for almost all modern approaches to edge detection (Canny, 1986) – and the well investigated functional regions of interest (ROI) in high level vision of human discovered using fMRI that consistently serve as face and place detectors (Kanwisher et al., 1997; Epstein et al., 1999). Outside of the field of vision, (Toneva & Wehbe, 2019) recently demonstrated that the explanatory arrow can be reversed in the domain of language, and that brain activity during reading can be used to facilitate the interpretation of deep neural network language models. In this light, we suggest that our extensive understanding of biological vision will not only enable future advances in artificial vision systems, but that this knowledge will also enable a better understanding of the inner workings of such systems.

2 METHODS

Encoding Models and Stacked Regression Encoding models (Naselaris et al., 2011) enable us to relate stimulus features and brain activity. If a feature is a good predictor of a specific brain region, information about that feature is likely encoded in that region. Here, we featurized each of stimulus images by extracting layerwise features from specific networks and use them in the voxelwise encoding models to predict brain responses in specific regions-of-interest (ROIs). All images are split into training and testing set. Model performances are reported as correlations between predicted responses and true responses.

To encode with multiple features, we applied the stacked regression method (Wolpert, 1992; Breiman, 1996). We adapted this approach such that each encoding model used a different feature space as input. At each voxel, encoding models are trained, then the stacking algorithm learns a convex combination of the predictions of these models for that voxel. The result from stacking is a readily-interpretable combination of individual features that outperforms the performance of the best feature alone. These stacking weights indicate how features are best combined to predict the specific voxel response: generally, the fewer errors a feature makes in its respective encoding model, the higher its corresponding stacked weight; that is, the importance of that feature for prediction.

Natural Scenes Dataset (NSD) NSD (Allen et al., 2021) is a large-scale fMRI dataset collected at ultra-high-field strength (7T). NSD consists of whole-brain, high-resolution measurements of 8 adult participants as they viewed thousands of color natural scenes over the course of 30–40 scan sessions. The natural scenes are obtained from the Microsoft COCO image dataset (Lin et al., 2014). Beta estimates are obtained from single-trial GLM in which the HRF is estimated for each voxel.

3 RESULTS

We extracted learned representations from different tasks and network architectures and explored how they differ in predicting brain responses to natural images. Layerwise features were extracted from specific networks and then used to build voxelwise encoding models for the cortical area of Participant 1 in NSD. For evaluation, we calculated R, the square root of the *coefficient of determination*, as the metric of the goodness of fit for the encoding model. We also show weights learned from the stacking algorithm for each feature.

To investigate how tasks influence the representations learned by a network and their ability to predict brain data, we fixed the network architecture and compared representations learned for object and scene classification. We used AlexNet (Krizhevsky et al., 2012) pretrained on ImageNet (Deng et al., 2009) and Places365 (Zhou et al., 2017) for each task. For each AlexNet model, we extracted features from the following 7 layers in an order consistent with the network architecture: Conv-1, Conv-2, Conv-3, Conv-4, Conv-5, FC-6, FC-7.

Figure 1 shows the result from encoding V1, V2, V3, V4 in the early visual cortex, Place ROIs (OPA, PPA, RSC) and Face ROIs (FFA, OFA, aTLface) using layerwise features from AlexNet for object and scene classification. Within each subfigure, each individual line is prediction performances and stacking weights across features from different layers for an individual voxel. For each row of the subfigures, we can see a progression of preferred layers across ROIs. Consistent with previous results (Yamins et al., 2014; Güçlü & van Gerven, 2015), in *AlexNet-Object* features extracted from convolution layers (Conv-2 and Conv-3) encode the early visual areas (especially V1, V2, V3) better

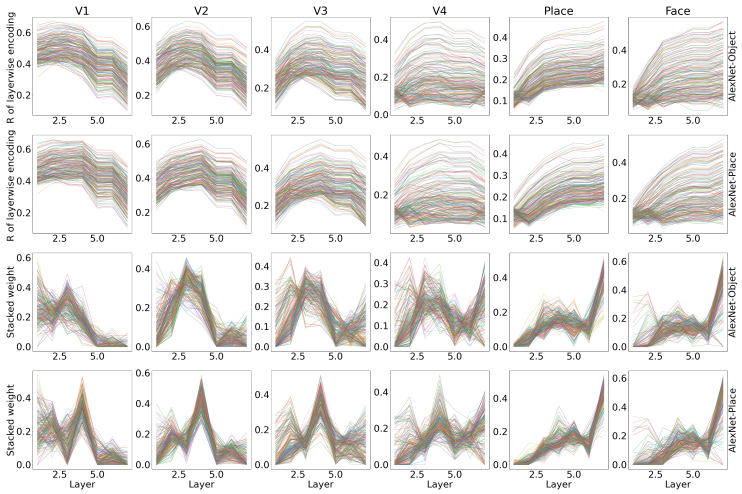

Figure 1: AlexNet encoding results. Every line corresponds to one of the best 200 encoded voxels. Each column corresponds to a visual ROI. The first and second rows are R results for *AlexNet-Object* and *AlexNet-Place* layers. The third and fourth row are the stacking weights. For V1-V4 show a reverse pattern when going from *AlexNet-Object* to *AlexNet-Place*: weights of Conv-4 layer surge and weights of Conv-3 layer plunge.

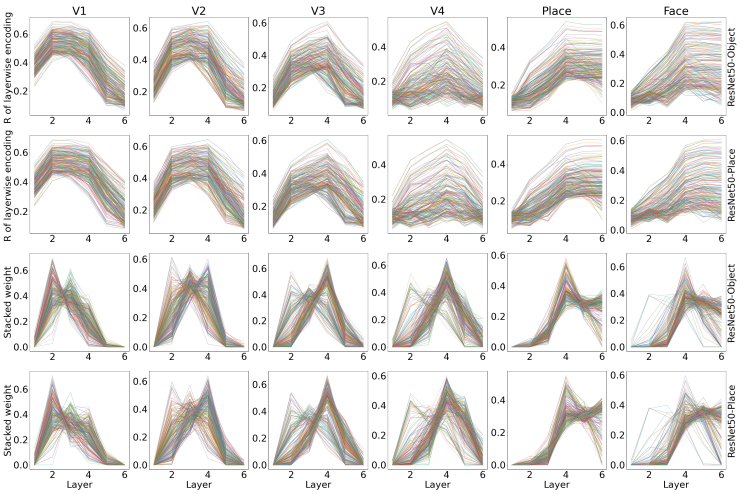

Figure 2: ResNet50 encoding results. Every line corresponds to one of the best 200 encoded voxels. Each column corresponds to a visual ROI. The first and second rows are R results for *ResNet50-Object* and *ResNet50-Place* layers. The third and fourth row are the stacking weights. Preferred layers for *ResNet50-Object* and *ResNet50-Place* are consistent across ROIs.

while features extracted from fully connected layer (FC-7) outperform those from all other layers in encoding Place and Face ROIs. Comparing the first row with the second, and the third row with the fourth, we can see there is a **peak weight shift** from Conv-3 layer to Conv-4 layer as we change the task from *AlexNet-Object* to *AlexNet-Place*. This indicates that with the same architecture, change of tasks could affect how representations from network predict the brain.

Task differences observed in Alexnet do not replicate when we change the network architecture to ResNet50 (He et al., 2016) while fixing the task and dataset. From ResNet50, we extract features from the following 6 layers in the order consistent with how the network is built: Conv-1, the last layer of Conv-2 to Conv-5 blocks respectively, and the last Avgpool layer before the final layer.

Figure 2 shows results from voxelwise encoding models of V1-V4 in the early visual cortex, Place ROIs (OPA, PPA, RSC) and Face ROIs (FFA, OFA, aTLface) using layerwise features from ResNet50 for object and scene classification. Similar to what we see in the AlexNet results in Figure1, we observe the same trend that features extracted from early layers represent the early visual cortex better while features extracted from later layers represent Place and Face ROIs better. However, preferred layers by the brain as well as stacking weight are consistent between the two tasks, indicating that the additional depth of networks might lessen the influence of task in terms brain prediction and that these deeper network might just represent more information about the input that are not subject to tasks. One thing to note is that, for Face and Place area prediction, layer 4 in ResNet is assigned the largest stacking weight. Different from what we see in Alexnet, where the last layer has the largest weight, it indicates that network expressiveness might the key for a good brain prediction instead of the semantic structure in the representations.

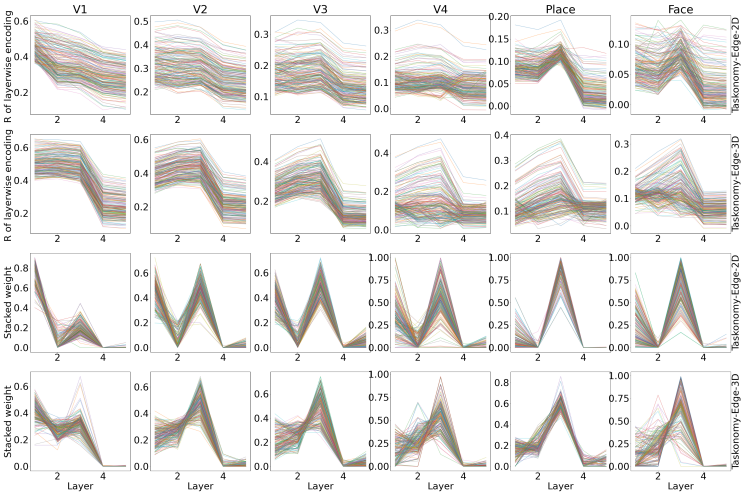

Figure 3: Taskonomy edge detection network encoding results. Content in each subfigure is similar as ones in previous figures. The first and second rows are R results for *Edge2D* and *Edge3D* layers. The third and fourth row are the stacking weights. For both *Edge2d* and *Edge3d*, early layers predict consistently better across ROIs and Conv-3 layer is the most preferred in all ROIs except V1.

Lastly, the commonly observed pattern that early layers in networks predict early visual layers in the brain better while later layers in a network predict higher visual areas better, as shown in Yamins et al. (2014); Güçlü & van Gerven (2015), does not hold when using representations extracted from edge detection networks. Here we extracted features from Taskonomy (Zamir et al., 2018) encoder trained for 2D and 3D edge detection. The network architecture is similar to ResNet50 but differs in replacing stride 2 convolution with stride 1 convolution in Conv-5 and removing all global average pooling. We extracted features exactly as what we do in ResNet50 but excluded Conv-4 blocks.

Figure 3 shows the encoding results for V1 to V4 in the early visual cortex, Place ROIs (OPA, PPA, RSC) and Face ROIs (FFA, OFA, aTLface) using Taskonomy features for 2D and 3D edge detection. For both *Edge2d* and *Edge3d*, early layers predict consistently better across ROIs. The overall prediction performance is lower, which is not surprising considering how little information edge detection task would normally required compared to more high level semantic tasks. What's surprising here is how early layers of edge detection networks yield higher prediction performances than the later layers even in predicting place and face areas in the brain. From the neuroscience literature, we know fairly well about the consistent responses of place and face images in Place (Epstein et al., 1999; Park & Chun, 2009; Rajimehr et al., 2011) and Face ROIs (Kanwisher et al., 1997; Tarr & Gauthier, 2000; Kanwisher et al., 2002; Gauthier et al., 2000; Grill-Spector et al., 2004) in the brain respectively. Contrary to the commonly believed view that a network trained to do a task should only represent variance relevant to that task (Bruna & Mallat, 2013), our result indicates that a network could possibly represent more information than what is needed in a task among the intermediate layers. Further analysis would be needed to further support this point.

4 Discussion

We observed that in predicting brain response using representations from a relatively simple neural network (i.e., AlexNet), varying the training task leads to differences in which network layers best predict the brain. Of note, this effect is network dependent and disappears when the same comparison is done with a much larger network (i.e., ResNet50). As previously shown in multi-task learning, network capacity and expressive power (Bengio & Delalleau, 2011; Raghu et al., 2017) influences the learned task-relevant representations and affects how different tasks may be learned together (Standley et al., 2020). Thus, our first takeaway is that network structure should be taken into consideration when mapping from network representations to the brain. A second takeaway is that, as exemplified by our results from edge detection networks, one can leverage our extensive understanding of the computations realized in different brain areas to gain a more holistic understanding of learned representations in neural networks - a step beyond visualizing randomly- or hand-picked units. Overall, the methods presented here enable a more comprehensive approach to using neural network representations to model brain function, allowing us to both better understand how choices as to network architecture and task affect predictions for biological systems and, conversely, to further interpret the learned representations realized in artificial systems.

ACKNOWLEDGMENTS

Thanks to Emily J. Allen, Yihan Wu, Thomas Naselaris and Kendrick Kay for collection and sharing the NSD dataset. Collection of the NSD dataset was supported by NSF IIS-1822683 and NSF IIS-1822929.

RESPONSE TO REVIEWERS

We are thankful for detailed feedback from both of our reviewers. Below we listed some specific responses to address their concerns and confusions.

R1: We have added more details in data preprocessing as well as model training in the main paper. For each result figure, we have also provided more of our interpretation.

R2: For this paper, we define task as a distinct input to output mapping. In this case, object and place classifications are considered different tasks because different images to labels mapping.

For results figures, we picked top 200 voxels only for visulization purpose. We would like to include as many voxels as possible while most of lines are still identifiable. We find 200 to be a good number for this purpose.

In terms of why intermediate features from an edge detection network can still predict areas like FFA and PPA, we agree with the reasoning that face and place areas do retain edge information while they processes these high level categories. Though this does not explain why we observe better prediction accuracy from the middle layers of edge detection networks instead of later layers. This reasoning, instead, would predict that all layers predict these areas similarly or the later layers predict better.

REFERENCES

Pulkit Agrawal, Dustin Stansbury, Jitendra Malik, and Jack L Gallant. Pixels to voxels: Modeling visual representation in the human brain. *arXiv preprint arXiv:1407.5104*, 2014.

Emily Jean Allen, Ghislain St-Yves, Yihan Wu, Jesse L Breedlove, Logan T Dowdle, Bradley Caron, Franco Pestilli, Ian Charest, J Benjamin Hutchinson, Thomas Naselaris, et al. A massive 7t fmri dataset to bridge cognitive and computational neuroscience. *bioRxiv*, 2021.

David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6541–6549, 2017.

Yoshua Bengio and Olivier Delalleau. On the expressive power of deep architectures. In *International conference on algorithmic learning theory*, pp. 18–36. Springer, 2011.

Leo Breiman. Stacked regressions. *Machine learning*, 24(1):49–64, 1996.

Joan Bruna and Stéphane Mallat. Invariant scattering convolution networks. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1872–1886, 2013.

J Canny. A computational approach to edge detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 8 (6):679–698, June 1986. ISSN 0162-8828. doi: 10.1109/TPAMI.1986.4767851. URL `https://doi.org/10.1109/TPAMI.1986.4767851`.

Charlotte Caucheteux and Jean-Rémi King. Language processing in brains and deep neural networks: computational convergence and its limits. *BioRxiv*, 2020.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Russell Epstein, Alison Harris, Damian Stanley, and Nancy Kanwisher. The parahippocampal place area: recognition, navigation, or encoding? *Neuron*, 23(1):115–125, 1999.

Isabel Gauthier, Michael J Tarr, Jill Moylan, Pawel Skudlarski, John C Gore, and Adam W Anderson. The fusiform "face area" is part of a network that processes faces at the individual level. *Journal of cognitive neuroscience*, 12(3):495–504, 2000.

Kalanit Grill-Spector, Nicholas Knouf, and Nancy Kanwisher. The fusiform face area subserves face perception, not generic within-category identification. *Nature neuroscience*, 7(5):555–562, 2004.

C G Gross. How inferior temporal cortex became a visual area. *Cerebral Cortex*, 4(5):455–469, 1994.

Umut Güçlü and Marcel AJ van Gerven. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, 35(27):10005–10014, 2015.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

D H Hubel and T N Wiesel. Receptive fields of single neurons in the cat's striate cortex. *J. Physiol.*, 148:574–591, 1959.

Shailee Jain and Alexander Huth. Incorporating context into language encoding models for fmri. In *Advances in neural information processing systems*, pp. 6628–6637, 2018.

Shailee Jain, Vy A Vo, Shivangi Mahto, Amanda LeBel, Javier S Turek, and Alexander G Huth. Interpretable multi-timescale models for predicting fmri responses to continuous natural speech. *Advances in Neural Information Processing Systems*, 2020.

Nancy Kanwisher, Josh McDermott, and Marvin M Chun. The fusiform face area: a module in human extrastriate cortex specialized for face perception. *Journal of neuroscience*, 17(11):4302–4311, 1997.

Nancy Kanwisher, Josh McDermott, and Marvin M Chun. The fusiform face area: A module in human extrastriate cortex specialized for face perception. In *Foundations in social neuroscience*. MIT Press Cambridge, MA, 2002.

Alexander JE Kell, Daniel LK Yamins, Erica N Shook, Sam V Norman-Haignere, and Josh H McDermott. A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. *Neuron*, 98(3):630–644, 2018.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Alexander Mordvintsev, Christopher Olah, and Mike Tyka. Inceptionism: Going deeper into neural networks, 2015. URL `https://research.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html`.

Thomas Naselaris, Kendrick N Kay, Shinji Nishimoto, and Jack L Gallant. Encoding and decoding in fmri. *Neuroimage*, 56(2):400–410, 2011.

Aran Nayebi, Javier Sagastuy-Brena, Daniel M Bear, Kohitij Kar, Jonas Kubilius, Surya Ganguli, David Sussillo, James J DiCarlo, and Daniel LK Yamins. Goal-driven recurrent neural network models of the ventral visual stream. *bioRxiv*, 2021.

Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. Feature visualization. *Distill*, 2017. doi: 10.23915/distill.00007. https://distill.pub/2017/feature-visualization.

Chris Olah, Arvind Satyanarayan, Ian Johnson, Shan Carter, Ludwig Schubert, Katherine Ye, and Alexander Mordvintsev. The building blocks of interpretability. *Distill*, 2018. doi: 10.23915/distill.00010. https://distill.pub/2018/building-blocks.

Soojin Park and Marvin M Chun. Different roles of the parahippocampal place area (ppa) and retrosplenial cortex (rsc) in panoramic scene perception. *Neuroimage*, 47(4):1747–1756, 2009.

Maithra Raghu, Ben Poole, Jon Kleinberg, Surya Ganguli, and Jascha Sohl-Dickstein. On the expressive power of deep neural networks. In *international conference on machine learning*, pp. 2847–2854. PMLR, 2017.

Reza Rajimehr, Kathryn J Devaney, Natalia Y Bilenko, Jeremy C Young, and Roger BH Tootell. The "parahippocampal place area" responds preferentially to high spatial frequencies in humans and monkeys. *PLoS Biol*, 9(4):e1000608, 2011.

Martin Schrimpf, Jonas Kubilius, Ha Hong, Najib J. Majaj, Rishi Rajalingham, Elias B. Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Kailyn Schmidt, Daniel L. K. Yamins, and James J. DiCarlo. Brain-score: Which artificial neural network for object recognition is most brain-like? *bioRxiv*, 2018. doi: 10.1101/407007.

Trevor Standley, Amir Zamir, Dawn Chen, Leonidas Guibas, Jitendra Malik, and Silvio Savarese. Which tasks should be learned together in multi-task learning? In *International Conference on Machine Learning*, pp. 9120–9132. PMLR, 2020.

Michael J Tarr and Isabel Gauthier. Ffa: a flexible fusiform area for subordinate-level visual processing automatized by expertise. *Nature neuroscience*, 3(8):764–769, 2000.

Mariya Toneva and Leila Wehbe. Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). In *Advances in Neural Information Processing Systems*, pp. 14928–14938, 2019.

Leila Wehbe, Ashish Vaswani, Kevin Knight, and Tom Mitchell. Aligning context-based statistical models of language with brain activity during reading. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014.

David H Wolpert. Stacked generalization. *Neural networks*, 5(2):241–259, 1992.

Daniel L K Yamins and James J DiCarlo. Using goal-driven deep learning models to understand sensory cortex. *Nat Neurosci*, 19(3):356–365, 2016. doi: 10.1038/nn.4244.

Daniel LK Yamins, Ha Hong, Charles F Cadieu, Ethan A Solomon, Darren Seibert, and James J DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the national academy of sciences*, 111(23):8619–8624, 2014.

Amir R Zamir, Alexander Sax, William Shen, Leonidas J Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3712–3722, 2018.

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.