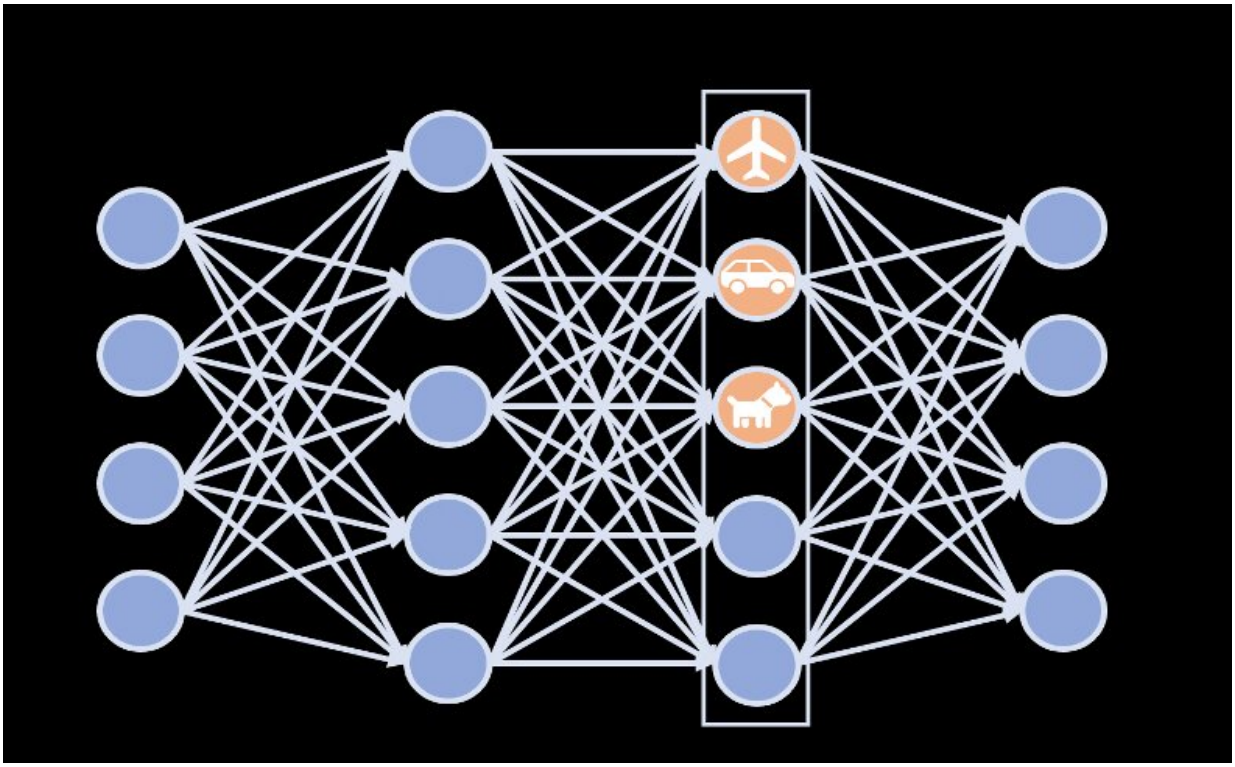
Concept whitening: A strategy to improve the interpretability of image recognition models

January 13 2021, by Ingrid Fadelli



Concept Whitening disentangles the latent space of the neural network so that its axes are aligned with predefined concepts, e.g., 'airplane', 'car' and 'dog'. This means all information about the concept gathered by the network up to that point travels through that concept's single node. Credit: Chen, Bei & Rudin.

Over the past decade or so, deep neural networks have achieved very promising results on a variety of tasks, including image recognition

tasks. Despite their advantages, these networks are very complex and sophisticated, which makes interpreting what they learned and determining the processes behind their predictions difficult or sometimes impossible. This lack of interpretability makes deep neural networks somewhat untrustworthy and unreliable.

Researchers from the Prediction Analysis Lab at Duke University, led by Professor Cynthia Rudin, have recently devised a technique that could improve the interpretability of deep neural networks. This approach, called concept whitening (CW), was first introduced in a paper published in *Nature Machine Intelligence*.

"Rather than conducting a post hoc analysis to see inside the hidden layers of NNs, we directly alter the NN to disentangle the latent space so that the axes are aligned with known concepts," Zhi Chen, one of the researchers who carried out the study, told Tech Xplore. "Such disentanglement can provide us with a much clearer understanding of how the network gradually learns concepts over layers. It also focuses all the information about one concept (e.g., "lamp," "bed," or "person") to go through only one neuron; this is what is meant by disentanglement."

Initially, the technique devised by Rudin and her colleagues disentangles the latent space of a neural network so that its axes are aligned with known concepts. Essentially, it performs a "whitening transformation," which resembles the way in which a signal is transformed into white noise. This transformation decorrelates the latent space. Subsequently, a rotation matrix strategically matches different concepts to axes without reversing this decorrelation.

"CW can be applied to any layer of a NN to gain interpretability without hurting the model's predictive performance," Rudin explained. "In that sense, we achieve interpretability with very little effort, and we don't lose accuracy over the black box."

The new approach can be used to increase the interpretability of deep neural networks for image recognition without affecting their performance and accuracy. Moreover, it does not require extensive computational power, which makes it easier to implement across a variety of models and using a broader range of devices.

"By looking along the axes at earlier layers of the network, we can also see how it creates abstractions of concepts," Chen said. "For instance, in the second layer, an airplane appears as a gray object on a blue background (which interestingly can include pictures of sea creatures). Neural networks don't have much expressive power in only the second layer, so it is interesting to understand how it expresses a complex concept like 'airplane' in that layer."

The concept could soon allow researchers in the field of deep learning to perform troubleshooting on the models they are developing and gain a better understanding of whether the processes behind a model's predictions can be trusted or not. Moreover, increasing the interpretability of deep neural networks could help to unveil possible issues with training datasets, allowing developers to fix these issues and further improve a model's reliability.

"In the future, instead of relying on predefined concepts, we plan to discover the concepts from the dataset, especially useful undefined concepts that are yet to be discovered," Chen added. "This would then allow us to explicitly represent these discovered concepts in the latent space of neural networks, in a disentangled way, to increase interpretability."

More information: Concept whitening for interpretable image recognition. *Nature Machine Intelligence*(2020). DOI: 10.1038/s42256-020-00265-z.

users.cs.duke.edu/~cynthia/lab.html