Deconfounded Representation Similarity for Comparison of Neural Networks

Tianyu Cui[1], Yogesh Kumar[1],
Pekka Marttinen[1], Samuel Kaski[1,2]

[1] Aalto University [2] The University of Manchester

BACKGROUND:

Representation similarities between NNs:

For input $X \in \mathbb{R}^{n \times p}$ and the corresponding m_1th and m_2th layer representations of NNs f_1 and f_2, $X_{f_1}^{m_1}$ and $X_{f_2}^{m_2}$, we

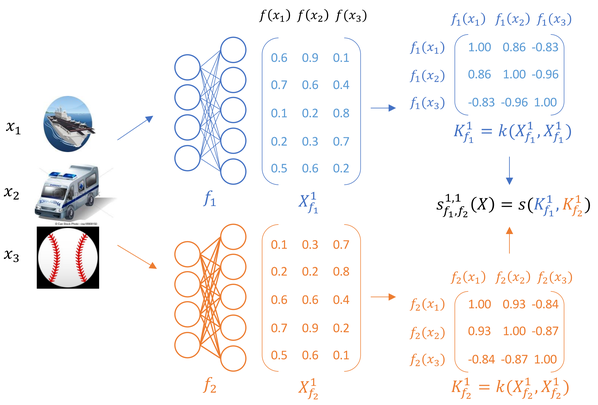
1. Compute the inter-example representation similarity matrices (RSMs) of $X_{f_1}^{m_1}$ and $X_{f_2}^{m_2}$:

$$K_{f_1}^{m_1} = k\left(X_{f_1}^{m_1}, X_{f_1}^{m_1}\right), K_{f_2}^{m_2} = k\left(X_{f_2}^{m_2}, X_{f_2}^{m_2}\right);$$

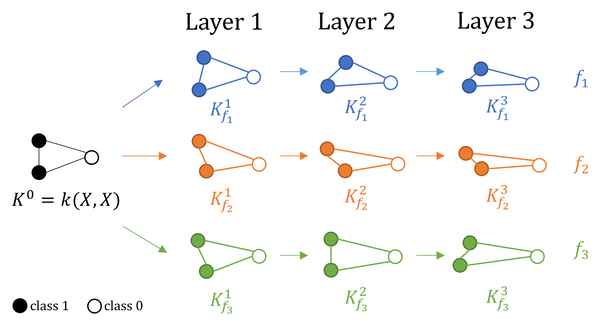
2. Compute the similarity between two RSMs:

$$s_{f_1,f_2}^{m_1,m_2} = s\left(K_{f_1}^{m_1}, K_{f_2}^{m_2}\right).$$

This covers most of methods, such as CKA[1].

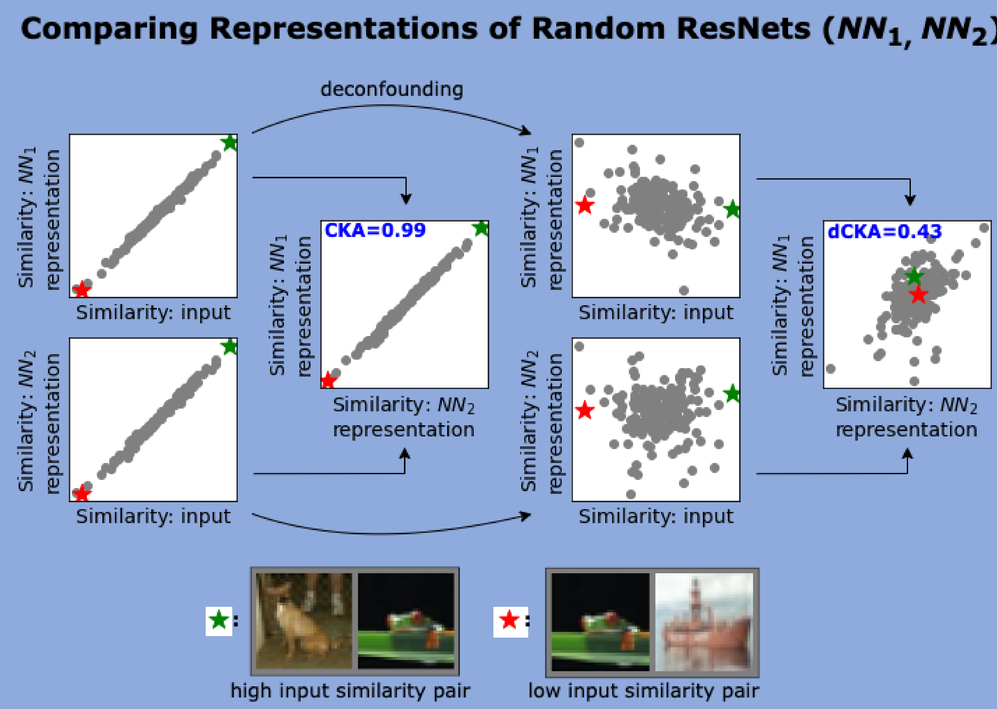


Neural networks modify RSMs layer-by-layer:



How to better compare representations learned by neural networks?

Current measures for comparing the similarity of representations between NNs may not well reflect their functional similarity due to the confounding effect of similarities of data items in the input space.



Counterintuitively, a conventional method (CKA[1]) gives completely random NNs a similarity close to one. Our dCKA gives a smaller similarity.

Take a picture to access the paper and code!

Selected as Oral

METHODS:

Deconfounded representation similarities:

Just regress out the confounder from RSMs:

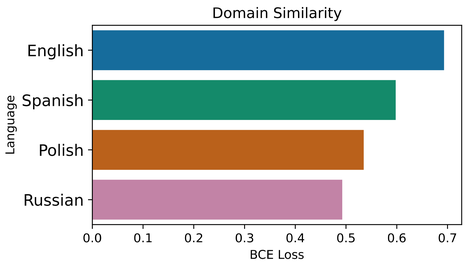
$$ds_{f_1,f_2}^{m_1,m_2} = s\left(K_{f_1}^{m_1} - \alpha_1 K^0, K_{f_2}^{m_2} - \alpha_2 K^0\right),$$

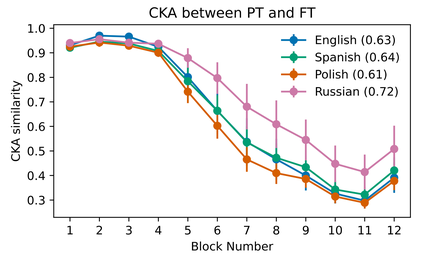
where α_i minimizes $\left\| K_{f_i}^{m_i} - \alpha_i K^0 \right\|_F$.

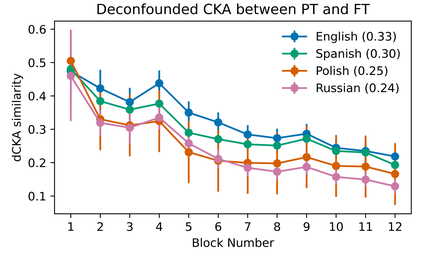
EXPERIMENTS:

Transfer Learning:

Domain similarity is inconsistent with CKA but consistent with the deconfounded CKA.







[1] Kornblith, Simon, et al. "Similarity of neural network representations revisited." *ICML*, 2019.

Aalto University
School of Science

The University of Manchester