Distilling Representational Similarity using Centered Kernel Alignment (CKA)

Aninda Saha[1,2], Alina Bialkowski[1], Sara Khalifa[2]

The University of Queensland[1], Data61 CSIRO[2]

We propose a novel loss function using one of the most robust neural network similarity metrics, Centered Kernel Alignment (CKA), to distil representational similarity, putting a greater emphasis on the shape rather than the scale of representational distribution.

Representation Distillation

Representation distillation is a promising knowledge distillation (KD) technique that exploits the structural knowledge of inter-class similarities to improve the performance of student models, as shown in Figure 1. Current state-of-the-art (SOTA) techniques apply distance-based metrics to distil the inter-example similarity matrices, which embody the representational similarity knowledge. However, these distance-based metrics are not invariant to isotropic scaling i.e. they penalise students for different scales of features, despite maintaining the correct shape of feature distribution, which ultimately dictates class separation.

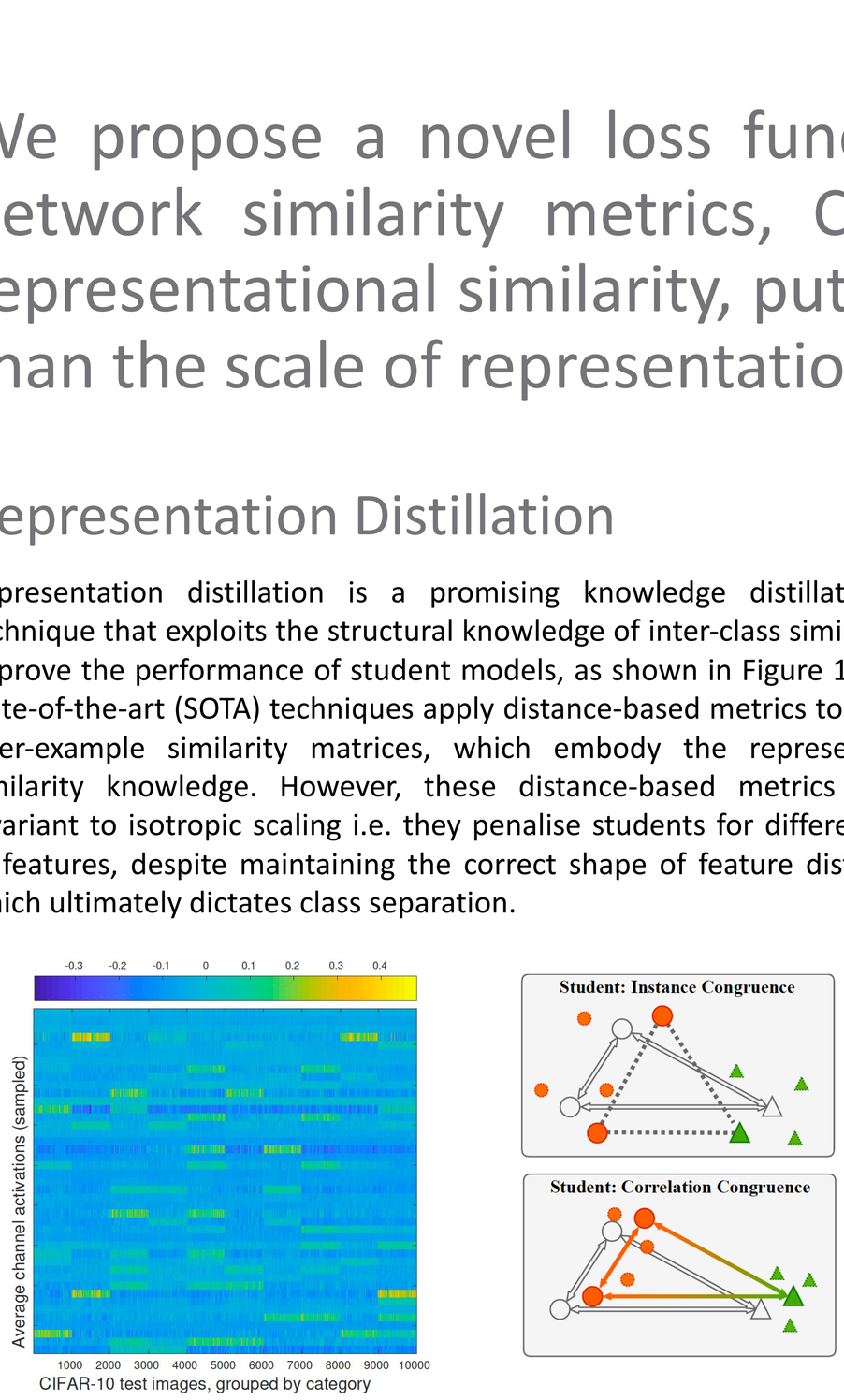

Figure 1: Tung et al. [1] (left) describes that semantically similar inputs produce similar activation patterns in a network, which indicates the presence of structural knowledge of class distributions. Peng et al. [2] (right) illustrates the difference between instance-level congruence, which is similar to feature-based distillation, and correlation congruence, which better captures the inter-class structural information and thereby improves student performance.

Centered Kernel Alignment (CKA)

Centered kernel alignment (CKA) is one of the most robust neural network similarity metrics [3], which applies centering and normalisation to inter-example similarity matrices using the Hilbert-Schmidt Independence Criterion (HSIC) kernel and then computes their cosine similarity.

$$CKA(G_S, G_T) = \frac{\mathrm{HSIC}(G_S, G_T)}{\sqrt{\mathrm{HSIC}(G_S, G_S) \cdot \mathrm{HSIC}(G_T, G_T)}}$$

This makes CKA invariant to isotropic scaling. Therefore, we propose a novel CKA-based loss for representation distillation as follows:

$$\mathcal{L}_{CKA} = 1 - CKA(G_S, G_T)$$

Through applying a loss metric that is invariant to isotropic scaling, we put a greater emphasis on shape rather than the scale of representational distribution, as shown in Figure 2. This is significant because the shape of the distribution dictates class separation, and therefore accuracy.

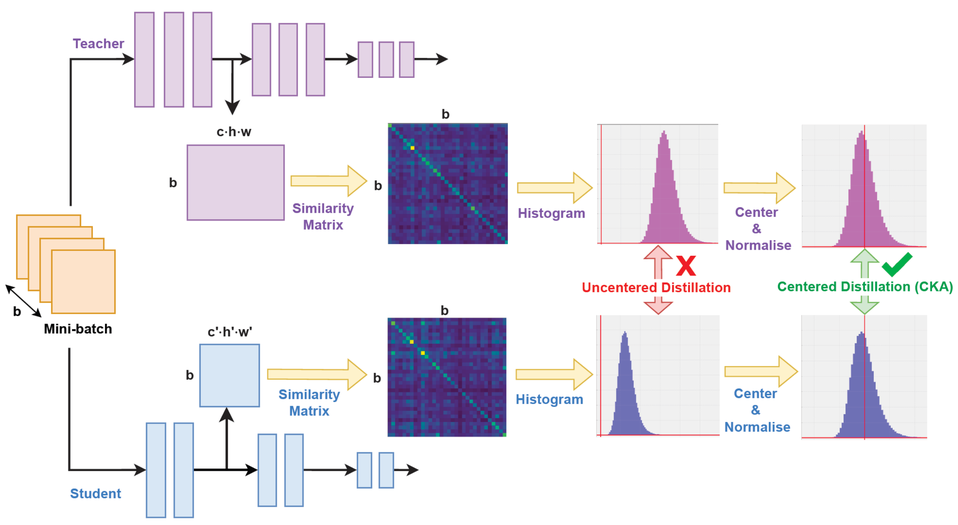

Figure 2: Our proposed framework for representation distillation computes the Gram matrices of intermediate feature maps, centers and normalises the matrices using the HSIC kernel and then distils them using the CKA loss metric. This allows us to put a greater emphasis on shape rather than the scale of representational distribution, which is more important for class separation, and therefore accuracy.

Results

We conducted a thorough evaluation of our hypothesis on CIFAR100, Tiny-ImageNet and ImageNet-1k datasets and compared our results with that of the three SOTA representation distillation methods: similarity-preserving (SP) KD, correlation congruence (CC) and relational KD (RKD). We also measured their performance on a ResNet18 and a MobileNetV2 student, to ensure generalisability across different datasets and architectures. The results for our CIFAR100 datasets are presented in Table 1 below.

Table 1: Results on CIFAR100 with two students ResNet18 (R18) and MobileNetV2 (MV2), distilled from a series of teachers using four main techniques: SP, CC, RKD and CKA (ours).

Architecture		Model Accuracy (%)					
Student	Teacher	Student	SP	CC	RKD	CKA	Teacher
R18	R18	77.98	78.22	78.54	78.63	**79.14**	77.98
R18	R34	77.98	78.41	78.94	78.83	**79.35**	78.97
R18	R50	77.98	78.15	78.46	78.32	**79.15**	79.19
R18	R101	77.98	78.27	78.69	78.47	**79.32**	79.68
MV2	R18	73.12	73.75	74.10	73.98	**75.30**	77.98
MV2	R34	73.12	73.47	74.04	73.89	**75.24**	78.97
MV2	R50	73.12	73.39	74.18	74.27	**75.19**	79.19
MV2	R101	73.12	73.32	74.14	74.00	**75.11**	79.68

Conclusion

We conclude from our study that the **cosine similarity** of the **centered** inter-example similarity matrices provides a less rigid yet accurate source of knowledge for the student to learn from, while being easier to apply than other forms of KD. The cosine similarity, being a loss between 0 and 1, provides a more stable training ground, leading to a better learning outcome for the student. We believe CKA has great potential for application in other vision problems such as object detection and segmentation as well.

FOR FURTHER INFORMATION

Aninda Saha

Email: a.saha@uqconnect.edu.au

REFERENCES

[1] Frederick Tung and Greg Mori. Similarity-preserving knowledge distillation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 1365– 1374, 2019.

[2] Baoyun Peng, Xiao Jin, Jiaheng Liu, Dongsheng Li, Yichao Wu, Yu Liu, Shunfeng Zhou, and Zhaoning Zhang. Correlation congruence for knowledge distillation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 5007–5016, 2019.

[3] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In International Conference on Machine Learning, pages 3519–3529. PMLR, 2019.

THE UNIVERSITY OF QUEENSLAND AUSTRALIA

BMVC 2022

CSIRO DATA 61