Speaker: Florentin GUTH

Title: On the universality of neural encodings in CNNs

Abstract: Deep networks achieve remarkable performance on many high-dimensional datasets, yet we cannot answer simple questions about what they have learned. For instance, do they learn the same "features" no matter their initialization? What about when we change the architecture or the training dataset? I will show how to meaningfully compare weights of deep networks using an alignment procedure on their hidden layers. We find that CNNs trained on image classification tasks share a common set of universal features, even for deep layers. These results explain, at a more fundamental level, the success of transfer learning, and pave the way for principled foundation models.