SHIKHAR JAISWAL, University of Toronto

Understanding The Modality Gap In Multi-Modal Systems

Multi-modal systems are integral components of machine learning models that process information from various data modalities, like text, images and audio by mapping these inputs into a shared representation space. Inherent to the development of these systems is the phenomenon of "modality gap", wherein, image embeddings and text embeddings are "distributed" differently within the shared representation space. This "gap" between the learnt representations has detrimental consequences for downstream tasks like search, retrieval and recommendation. While several empirical studies have attributed this phenomenon to the inherent difference in the distributional richness of the modalities, the mathematical analysis of its origins remains limited. In this talk, we will (i) define the modality gap in a generalized way, with a focus on its effect for downstream applications; (ii) present recent studies to better understand and mitigate this issue; and (iii) analyze the problem under simplifying assumptions regarding the network architecture and data distribution.