Contribution ID: **34** Type: **Oral presentation**

Semi-supervised multimodal representation learning through a global workspace

Tuesday 1 October 2024 14:15 (30 minutes)

Recent deep learning models can efficiently combine inputs from different modalities (e.g., images and text) and learn to align their latent representations, or to translate signals from one domain to another (as in image captioning, or text-to-image generation). However, current approaches mainly rely on brute-force supervised training over large multimodal datasets. In contrast, humans (and other animals) can learn useful multimodal representations from only sparse experience with matched cross-modal data. Here we evaluate the capabilities of a neural network architecture inspired by the cognitive notion of a "Global Workspace": a shared representation for two (or more) input modalities. Each modality is processed by a specialized system (pretrained on unimodal data, and subsequently frozen). The corresponding latent representations are then encoded to and decoded from a single shared workspace. Importantly, this architecture is amenable to self-supervised training via cycle-consistency: encoding-decoding sequences should approximate the identity function. For various pairings of vision-language modalities and across two datasets of varying complexity, we show that such an architecture can be trained to align and translate between two modalities with very little need for matched data (from 4 to 7 times less than a fully supervised approach). The global workspace representation can be used advantageously for downstream classification and cross-modal retrieval tasks and for robust transfer learning. Ablation studies reveal that both the shared workspace and the self-supervised cycle-consistency training are critical to the system's performance.

Contribution length

Middle

Primary authors: DEVILLERS, Benjamin (Centre de Recherche Cerveau et Cognition (CerCo), CNRS); MAYTIÉ, Léopold (Université Toulouse III - Paul Sabatier, Toulouse, France & Artificial and Natural Intelligence Toulouse Institute (ANITI)); VANRULLEN, Rufin (Centre de Recherche Cerveau et Cognition (CerCo), Artificial and Natural Intelligence Toulouse Institute (ANITI))

Presenter: MAYTIÉ, Léopold (Université Toulouse III - Paul Sabatier, Toulouse, France & Artificial and Natural Intelligence Toulouse Institute (ANITI))