Abstract

In recent years, there have been many attempts of decoding computational representations of sentences from the associated brain activity of human's reading these sentences during functional Magnetic Resonance Imaging. However, most of these studies have not interpreted what information is encoded in the representations, while assuming that these representations capture general constructs such as "semantics". One goal of this thesis is to demonstrate how computational representations can be characterized by how much information they contain related to many linguistic and semantic properties including animacy, semantic similarity, or word order to show they encode more than just the task they were trained on (e.g., text categorization). Therefore, if certain regions decode these representations, these regions can be better described. A second goal is to show the value of using complementary computational representations of the same task for brain mapping. For instance, different models –from logistic regression trained on tf-idf features to a deep neural network trained on word embeddings– accomplish text categorization in different ways. It is then possible to see which which brain regions have information on all forms of text categorization and which brain regions are specific to each model.

We first train models on text categorization. We then show how these models can be mapped onto the brain using Representational Similarity Analysis. However, in order to understand what information is mapped to each region, we provide a series of methods to analyze representations. We first compare each model's performance with human judgments on semantic textual similarity and text categorization to understand which model best captures human processing. The human sentence categorization task is carried out on sentences with more than one plausible category, which creates a challenge for any computational model that wishes to capture human intuition. Finally, we use a series of techniques to interpret each model's representations including probing, distorting inputs and architecture, Representational Similarity Analysis, and measuring how superficial or abstract each hidden layer is. We expect models that are similar in their interpretations to produce similar brain mappings.

Keywords: semantics, categorization, compositionality, convolutional neural network, long short-term memory, deep learning, brain, fmri.