Additive Analogies Reveal Compositional Structure in Neural Network Weights

Abi Tenenbaum
Yale University, New Haven, Connecticut, United States

R. Thomas McCoy
Yale University, New Haven, Connecticut, United States

Abstract

A central question in cognitive science is how to reconcile connectionist and symbolic models of the mind (e.g., Fodor & Pylyshyn 1988, Smolensky & Legendre 2006). Attempts have been made to bridge these competing schools of thought by showing how compositional structure can emerge in continuous vector representations (e.g., Manning et al. 2020). A key example is Mikolov et al. (2013), who demonstrated that word embeddings learned by a neural network encode semantic structure: subtracting the vector "man" from "king" and adding "woman" approximates "queen" (i.e., king - man + woman ≈ queen). Our work moves up one level of abstraction, from representations to functions. We analyze whether entire networks display emergent compositional structure by treating a trained network as a single vector (obtained by concatenating the network's parameters) encoding its function. We show that these parameter vectors can be recomposed through simple additive analogies to create networks with new functions.

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*