ON THE PITFALLS OF ANALYZING INDIVIDUAL NEURONS IN LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

While many studies have shown that linguistic information is encoded in hidden word representations, few have studied individual neurons, to show how and in which neurons it is encoded. Among these, the common approach is to use an external probe to rank neurons according to their relevance to some linguistic attribute, and to evaluate the obtained ranking using the same probe that produced it. We show two pitfalls in this methodology: 1. It confounds distinct factors: probe quality and ranking quality. We separate them and draw conclusions on each. 2. It focuses on encoded information, rather than information that is used by the model. We show that these are not the same. We compare two recent ranking methods and a simple one we introduce, and evaluate them with regard to both of these aspects.