White-box Testing of NLP models with Mask Neuron Coverage

Arshdeep Sekhon, Matthew Dwyer, Yangfeng Ji and Yanjun Qi, (NAACL 2022)

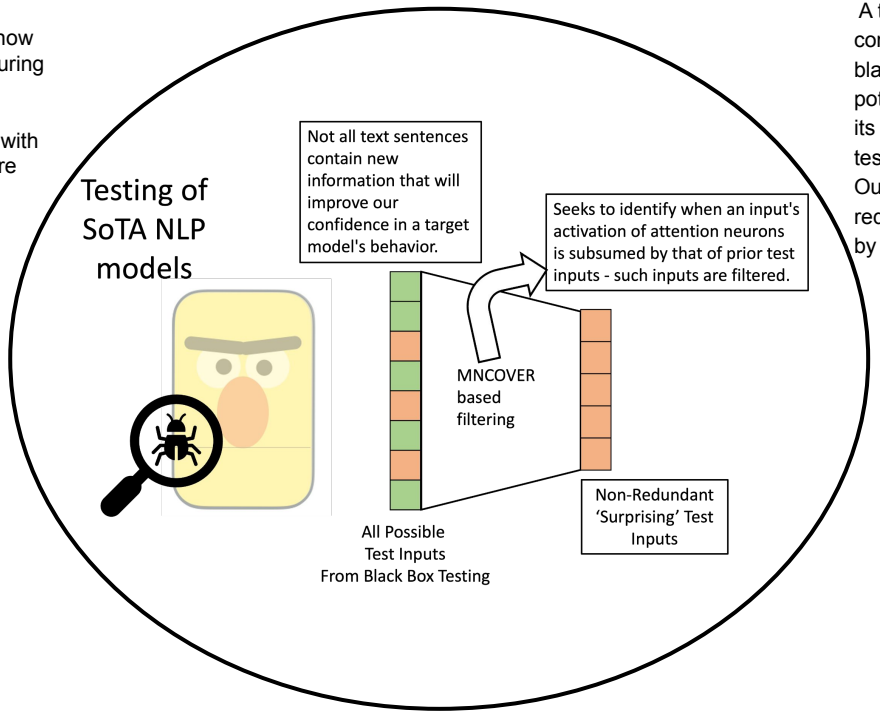
UNIVERSITY of VIRGINIA

Challenge

State of the art NLP models are software now being deployed in real world settings. Ensuring reliable behavior for these models is of paramount concern. However, these are notoriously black boxes and huge models with billions of parameters: off-the-shelf software testing methods are not applicable.

Solution

- ★ White box testing methods tailored to NLP models to thoroughly check their internal behavior. These can be used in conjunction with well established black box testing methods to ensure comprehensive testing.
- ★ A test coverage metric designed to address characteristics of NLP models and to account for the data distribution by considering task-specific important words and combinations.
- ★ Can substantially reduce the size of test sets while improving the failure detection of the resulting tests.



Scientific Impact

A tractable new coverage criteria enables comprehensive testing of NLP model both from a black-box and a white-box perspective. This has the potential to reduce the cost of testing without reducing its error-detection effectiveness, enabling more thorough testing.

Our proposed method provides a general framework to reduce cost of coverage based testing for NLP models by learning relevant neurons.

Broader Impact and Broader Participation

As NLP models get deployed across many application domains and industries, there is a pressing need to formalize how ML models get built, deployed, and behave. When real users start using it, the behavior could be completely different. Our proposed method enables comprehensive testing by complementing white box testing with black box testing of NLP models. This has consequences with respect to model trustworthiness and reliability.

Project Details: