Stability Analyses

David Spearman

November 2022

1 Introduction and Methodology

This brief discusses two hypotheses regarding the Stability AI memorization natural language processing project. The first hypothesis is that the order in which tokens are presented to a language prediction model for training is uncorrelated with the likelihood of memorization occurring successfully. The second hypothesis is that, for two different models A and B, the likelihood of a given token being memorized in A is unrelated with the likelihood of it being memorized in B.

For the first hypothesis, I ran a Person median skewness test on the accuracy of memorization, with the presentation order index as the axis of interest. Median skewness is appropriate rather than modal skewness because the data presented as multimodal, and therefore a modal skewness test would likely yield unreliable results. I employed the scipi.stats.skew method which uses the function $K = 3\frac{\bar{X}-M}{s}$ where K is skewness, $\bar{X}$ is the mean of the variable of interes, M is the median value of the variable of interest with respect to the axis, and s is the standard deviation of the variable of interest.

If this test yields results close to 0, then this indicates either that the hypothesis can be accepted or that the data most likely to be memorized are in the middle of the dataset. If this result is returned, a kurtosis test can resolve the ambiguity, with a kurtosis value of less than -1 indicating that the hypothesis can be accepted and a kurtosis value significantly higher than -1 indicating that tokens in the middle of the dataset are more likely to be memorized.

If the skewness test yields significantly positive results, that indicates that the hypothesis should be rejected and indicates that tokens which show up earlier in training are more likely to be memorized. Similarly, a significantly negative result indicates that the hypothesis should be rejected and that tokens later in training are more likely to be memorized.

For the second hypothesis, I used an ordinary least squares regression of the accuracy rate across observations in model A versus model B for all provided models. Strictly speaking, this is not a direct test of the hypothesis as accuracy data is a proportion value of the fraction of tokens which were successfully memorized, not token-by-token dummy variables. However, since token-by-token memorization similarity provides correlation on the observation-level, this

indicates that the methodology used may be a workable instrumental variable for an OLS regression.

In order to satisfy the assumptions of instrumental variables regressions, however, the token-by-token correlation must be the only way for observation-level accuracy to correlate. I can think of no other reason why this might be the case, but would advise caution until this assumption is confirmed by someone more familiar with the models in question.

If the assumption specified above holds, then this regression will produce insignificant results will be confirmatory for the hypothesis. A significantly positive result will indicate cross-model memorization is correlated, and therefore that a token in any given model A is likely to be memorized in model B for any given values of A and B. A significantly negative result will indicate anti-correlation, that a token is less likely to be memorized in B if it is memorized in A.

As this test requires running multiple regressions on various connected-but-distinct datasets, a Bayesian aggregation method may be necessary to compile results if there is sufficient conflict between them to make conclusions non-obvious.

2 Results

[FILL THIS IN WHEN YOU HAVE RESULTS]

3 Discussion

[FILL THIS IN WHEN THERE'S SOMETHING TO DISCUSS]