Final Project Checkpoint

Soham Patel and Yash

November 23rd, 2021

For our final project, we will be analyzing the *ELECTRA-small*'s performance on the Stanford NLI dataset (SNLI). As for our analysis, we'd like to look into using contrast/adversarial examples and try to also change the model in hopes of understanding which change would help improve the performance more (in terms of accuracy improvements on the SNLI). While the SNLI dataset doesn't contain the same range of genres as the Multi NLI dataset does, we decided that given the various pretrained *ELECTRA-small* weights using the SNLI (on huggingface), we would have significantly more benchmarks to compare to without having to spend several hours on training on colab or our personal machines.

Although the SNLI dataset contains roughly $570K$ sentence-hypothesis pairs, we realized that it would take extremely long to constantly retrain through all the examples while modifying the dataset (adding constrast or adversial examples). To save time, we decided to train on a significantly smaller subset of pairs ($10,000$ pairs) to reduce the training time to an hour and acquire a benchmark. We reasoned that given the lack of diverse genres in the dataset, the benchmark shouldn't vary significantly given an entire new subset of $10,000$ pairs. As mentioned on the README as well as corroborated by other huggingface benchmarks, the *ELECTRA-small*'s accuracy after training on the full SNLI dataset is 89% roughly (after 3 epochs) and after training on $10,000$ pairs with a batch size of 8, the *ELECTRA-small*'s accuracy on the evaluation set is roughly 43% (after 3 epochs). With this benchmark established on a subset of the data, we can use it to periodically compare any performance improvements when adding constrast examples (through manual construction) or adversarial challenges (through the TextAttack framework).

For generating the adversarial examples, we are currently looking into using the TextAttack Python Framework as a starting point, specifically exploring the TextFooler feature to rpelace sentences with their synonyms and observe the accuracy changes (if any) after replacing words. Regardless of the word changing which in turn will singificantly modify the inputs/outputs of the generator and discriminator, we speculate that the accuracy **shouldn't** increase or decrease that much given that the SNLI dataset should be fairly generalized. After modifying some examples from the evaluation set, the *ELECTRA-small*'s accuracy was reduced from 43% to 35%, which was significantly larger difference than we expected. From our understanding of NLI and the TextFooler feature, while based on the dictionary the words were replaced with synonyms, the replacements could lead to grammar errors as well as loss of comprehension (similar to machine translations as in certain phrases, synonyms cannot be swapped purely due to the context). With this accuracy drop on 10000 pairs, we speculate that the fully trained model might also experience a similar accuracy dorp (not as large due to generalization with more training points) and will have to look into whether this is due to the TextFooler or a non-robsut model.

We started with adversarial examples due to the framework provided as it would be fairly simply to generate adversarial examples and allow us to focus on the analysis. Along with investigating the true cause of the accuracy drop, we plan to add in contrast sets (fairly small if done manually) to determine which method works best in terms of accuracy increase with data modifications. We plan to have this analysis completed by Monday (November 29th) and begin implemeting potential fixes (which will most likely only entail adjusting the training set).