Adversarial evaluation of model performances

Here is an example on evaluating a model using adversarial evaluation of natural language inference with the Heuristic Analysis for NLI Systems (HANS) dataset McCoy et al., 2019. The example was gracefully provided by Nafise Sadat Moosavi.

The HANS dataset can be downloaded from this location.

This is an example of using test_hans.py:

```
export HANS_DIR=path-to-hans
export MODEL_TYPE=type-of-the-model-e.g.-bert-roberta-xlnet-etc
export MODEL_PATH=path-to-the-model-directory-that-is-trained-on-NLI-e.g.-by-using-
run_glue.py

python run_hans.py \
        --task_name hans \
        --model_type $MODEL_TYPE \
        --do_eval \
        --data_dir $HANS_DIR \
        --model_name_or_path $MODEL_PATH \
        --max_seq_length 128 \
        --output_dir $MODEL_PATH \
```

This will create the hans_predictions.txt file in MODEL_PATH, which can then be evaluated using hans/evaluate_heur_output.py from the HANS dataset.

The results of the BERT-base model that is trained on MNLI using batch size 8 and the random seed 42 on the HANS dataset is as follows:

```
Heuristic entailed results:
lexical_overlap: 0.9702
subsequence: 0.9942
constituent: 0.9962

Heuristic non-entailed results:
lexical_overlap: 0.199
subsequence: 0.0396
constituent: 0.118
```