Assignment

NLP Course

April 13, 2024

1 Introduction

You are tasked with training a transformer based model to answer multiple choice questions which are based on commonsense reasoning. The data set to be used for this task is SWAG : Situations With Adversarial Generations.

1.0.1 About Dataset

It consists of a question, the expected answer and upto 4 endings which include incorrect answers called distractors. The task is to train a model capable of identifying the correct answer among distractors. More details on the dataset can be found in the link provided.

2 Tasks

1. Read the dataset from the link provided and understand the task, Preprocess the data - choose/filter out the columns of interest. Use only the train split of the dataset for training the model. Ensure the model is not trained with the validation or test data.
2. Formulate the input to have the expected answer along with any 3 distractors (endings)
3. Define the tokenizer and use it to prepare the data as input to the model
4. You are free to choose a model architecture of your choice. Starting with pretrained models is allowed. Usage of models trained specifically for this task is discouraged.
5. Models can be customized as desired. For example adding modifications to layers, changing architecture, etc.
6. The expected output for this classification task is predicting the index corresponding to correct answer.
7. For evaluation Accuracy, Precision, Recall metrics can be used

3 Resources

- Source data
- Hugging face version of data
- Paper on the SWAG dataset
- Relevant papers that use this dataset