Investigating the Impact of Multi-Task Learning strategies on Selective Prediction

Arut Selvan Dhanasekaran	ASU ID: 1222275847	adhanas4@asu.edu
Sai Surya Kaushik Punyamurthula	ASU ID: 1220096111	spunyam2@asu.edu
Nikhil Chandra Nirukonda	ASU ID: 1223333995	nnirukon@asu.edu
Divya Reddy Katpally	ASU ID: 1222889889	dkatpall@asu.edu
Nikhitha Munugala	ASU ID: 1222913536	nmunugal@asu.edu

Motivation

The idea of Selective Prediction is to avoid making wrong predictions as the cost of making an incorrect prediction is much higher than refusing to make any prediction in some cases. (eg. Medical predictions). A model can be calibrated to be selective and reject a prediction if it's not confident enough. It should also be noted that the classifier should not reject all the instances due to lack of confidence.

Multi-Task Learning is the process of learning multiple tasks that are different from each other and creating a model that can solve multiple tasks instead of a model trained to solve only one specific task. Multi-task learning provides a way to perform better in a task with less data by leveraging the information from a related task. However, in general the multi-task models perform less as opposed to respective single-task models.

Objective

The goal of our project is to train a model on various datasets and study the impact of different multi-task learning strategies on the model's selective prediction performance.

Sampling Strategies used

- ***Heterogeneous Sampling*** - combine the training datasets of all tasks, shuffle them, and then train the model using this combined dataset.
- ***Homogeneous Sampling*** - combine the dataset but in one batch give examples of only one task and the batches can be shuffled

Models and Datasets

Models used

- BERT-base-cased
- T5-base (For experiment)

Datasets used

- SNLI - https://nlp.stanford.edu/projects/snli/
- SWAG - https://arxiv.org/abs/1808.05326
- Commonsense QA - https://www.tau-nlp.org/commonsenseqa
- Abductive NLI - https://arxiv.org/abs/1908.05739
- Social IQA - https://leaderboard.allenai.org/socialiqa/submissions/get-started

Experiments Performed

Multi-Task Learning using T5

Before we started working with BERT, we wanted to try out T5 first to check it's feasibility of training, training time and performance. For that first, we converted all the datasets to two sentences format where ***sentence1*** contains all the input data and ***sentence2*** contains the expected text answer. (Refer Fig 1). We fine tuned the ***T5 base*** model using ***Heterogeneous Sampling*** method (8000 samples from each dataset) and found it's accuracy to be better than that of BERT which is explained in the next section. But since the training took too long for T5 and lack of resources to train T5 faster, we decided to proceed with BERT QA. The training parameters and results for the above experiment are provided below.

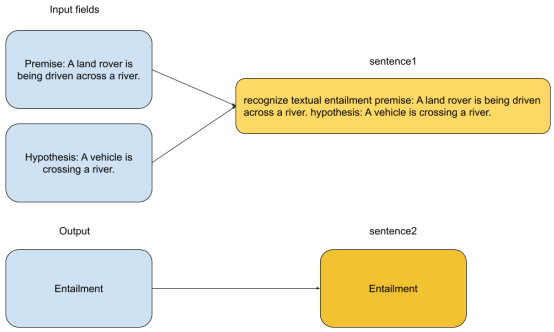

Fig 1: Converting a SNLI dataset instance to sentence1 and sentence2 for T5 training

T5 Training Params:

- Model: t5-base
- Batch Size (Training and Evaluation): 4
- Gradient Accumulation Size: 4
- Maximum Input Sequence Length - 175
- Adam Epsilon: 1.0e-08
- Epochs Trained: 5

T5 Accuracy (Exact match - Heterogeneous Sampling):

- SNLI - 77.67 %
- SWAG - 42.9 %
- Commonsense QA - 37.46 %
- Abductive NLI - 78.0 %
- Social IQA - 63.47 %

Multi-Task Learning using BERT

We changed our idea and decided to convert our datasets to "***question, context , answer***" format similar to SQUAD. The datasets are converted such that the context always contains the answer - a span of text present in the context (Ref Fig 2).


```
{
    "id":"swag_0",
    "question":"Sentence 1: Now, as someone coils a handful of rope, he glances at someone who stares off with a stunned gaze, his head propped on his hand. Sentence 2: Someone",
    "context":"Endings: smiles, then walks away with a frown., flows from her knuckles to lead ahead., shoots someone a disbelieving glance., crumples the cement assistance",
    "answer":{"answer_start":[88],"text":["shoots someone a disbelieving glance"]},
    "dataset":"swag"
}
```


Fig 2: Example instance of converted SWAG dataset

After the conversion, we trained a BERT-base-cased model using a heavily modified version of Huggingface QA scripts to support two different data sampling methods mentioned above - ***Homogenous Sampling*** and ***Heterogeneous Sampling*** *of the combined dataset*. For Homogenous Sampling, We wrote a custom sampler to homogeneously sample the combined dataset (8000 samples from each dataset) such that every batch contains only one type of dataset instances. The other training parameters remained the same for both the sampling strategies. The training parameters used are as follows.

BERT Training Params:

- Model: bert-base-cased
- Model Architecture: BertForQuestionAnswering
- Max Sequence Length: 256
- Training Batch Size: 16
- Gradient Accumulation Size: 1
- Maximum Input Sequence Length - 175
- Epochs Trained: 5
- Dropout Probability: 0.1
- Layer Normalization Epsilon: 1e-12

The two different models obtained after training using the above sampling methods are validated using the unseen samples (validation split) of the above datasets. The accuracy of the models are as below.

BERT Accuracy (Exact match):

	Homogenous Sampling	**Heterogeneous Sampling**
SNLI	76.86	79.62
SWAG	56.78	56.9
Commonsense QA	40.38	39.31
Abductive NLI	53.72	54.57
Social IQA	49.03	48.41

Except for Social IQA and Commonsense QA, the Heterogeneous sampling strategy slightly edged the Homogeneously trained model. Each prediction generated with the validation samples contains the best prediction with it's probability score. Those predictions are then fed into another script along with the validation dataset to compute ***Exact match***. Then the exact match, maximum probability are passed to another script to plot selective prediction metrics.

Selective Prediction Metrics

We use a few more metrics apart from Accuracy to measure selective prediction performance.

- ***Coverage***: For given threshold, the probability that model predicts a label instead of refusing making prediction
- ***Error***: For a given threshold, the probability that the true answer is different from the model's prediction when the model makes a prediction.
- ***Risk***: $\frac{Error}{Coverage}$

Selective Prediction Metrics Plots (Results)

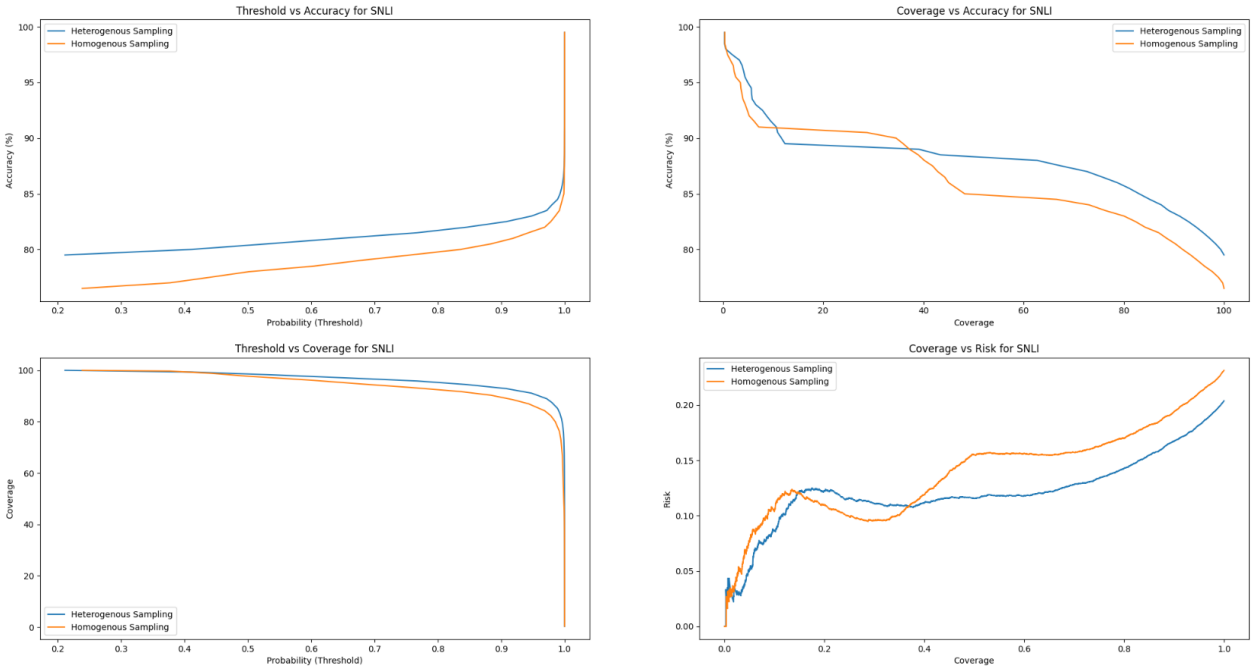

Fig 3: Selective Prediction Metrics Plots for SNLI (Homogenous and Heterogenous Sampling)

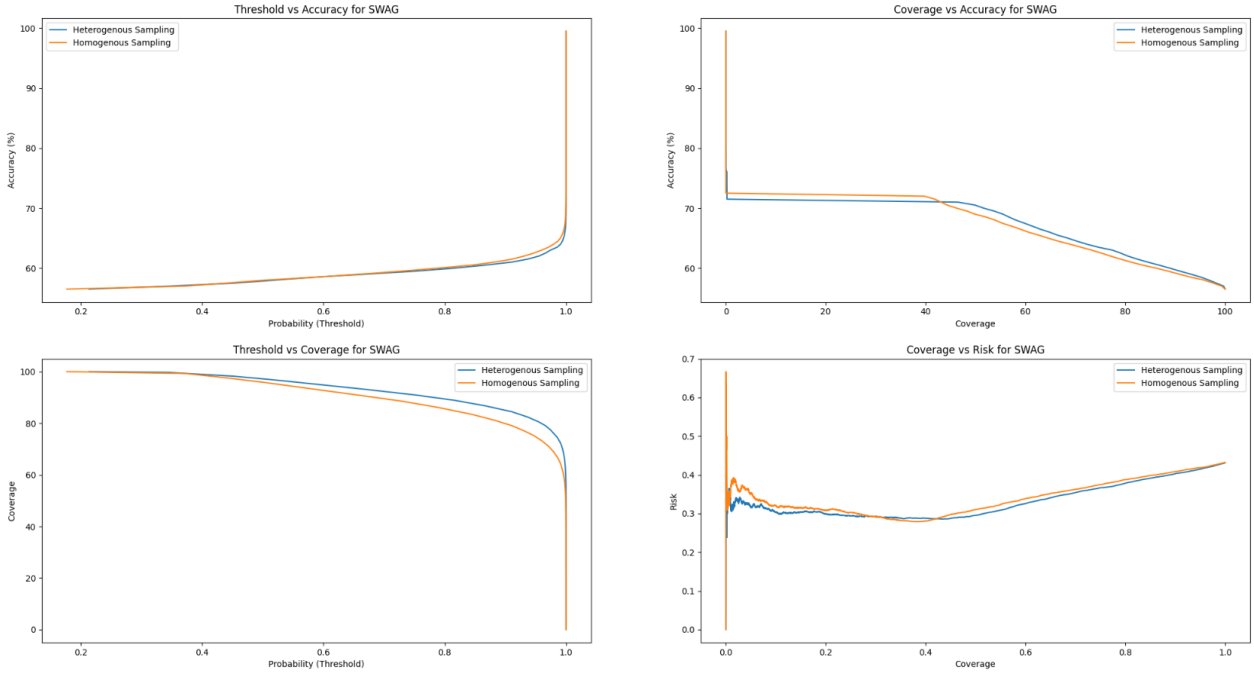

Fig 4: Selective Prediction Metrics Plots for SWAG (Homogenous and Heterogenous Sampling)

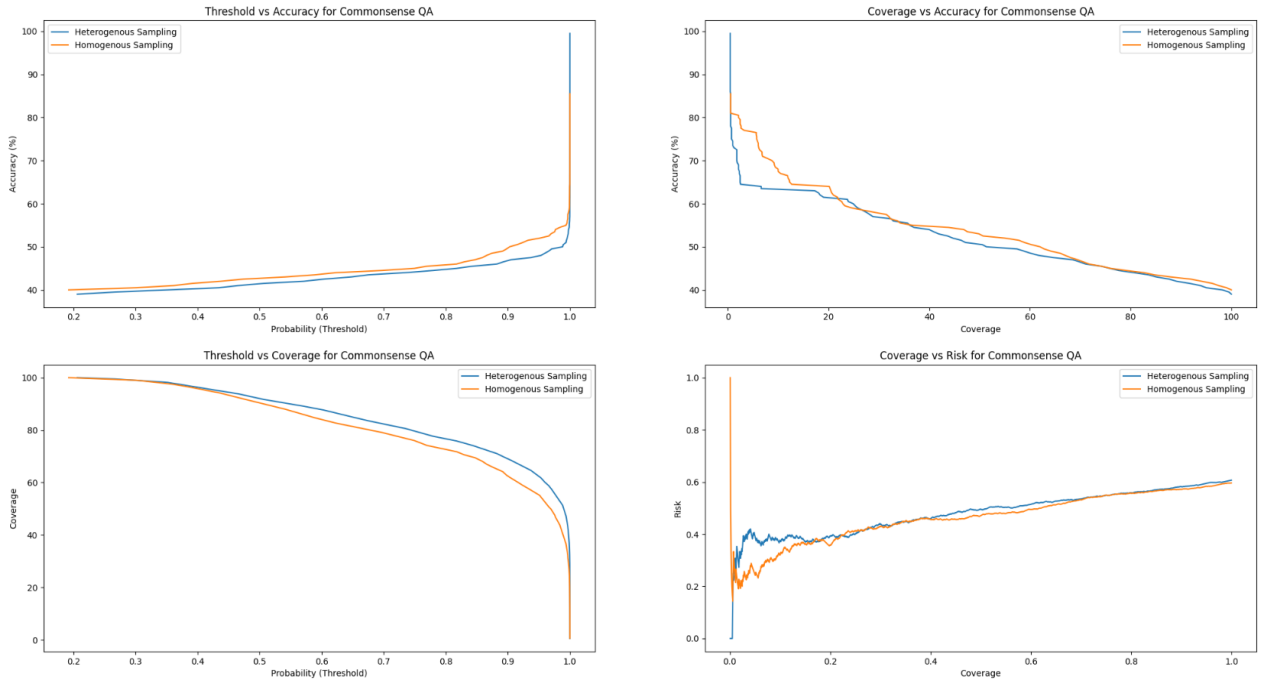

Fig 5: Selective Prediction Metrics Plots for Commonsense QA (Homogenous and Heterogenous Sampling)

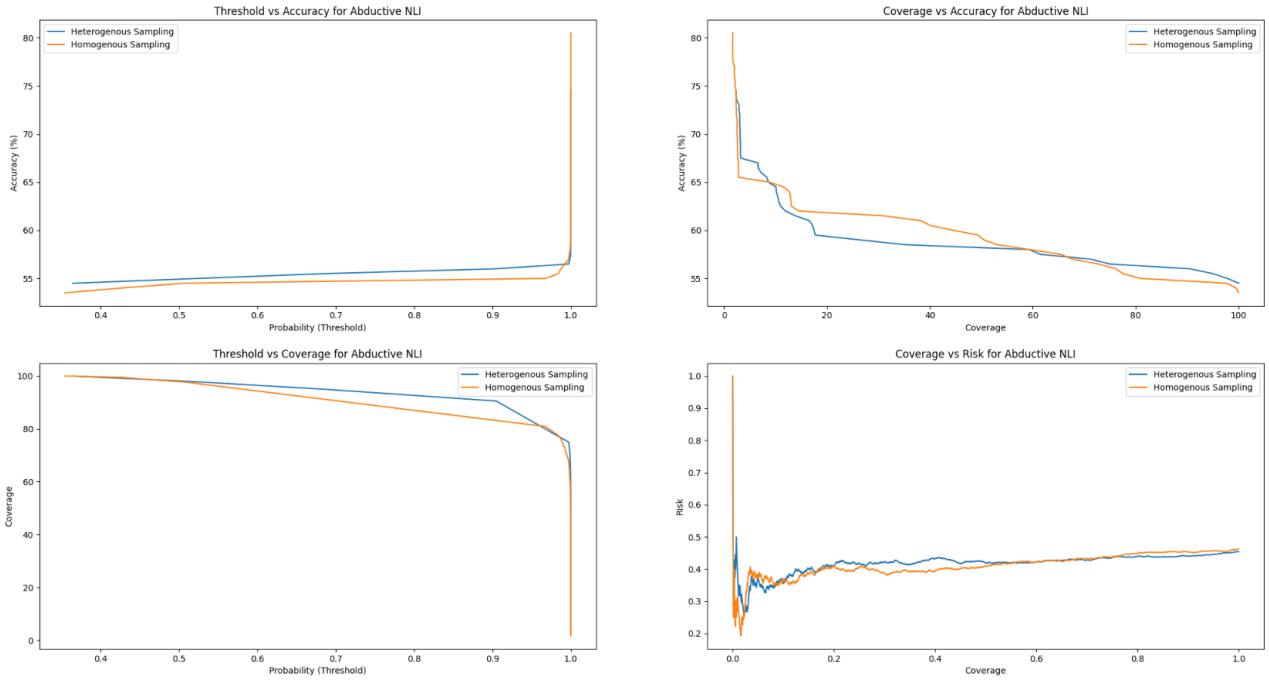

Fig 6: Selective Prediction Metrics Plots for Abductive NLI (Homogenous and Heterogenous Sampling)

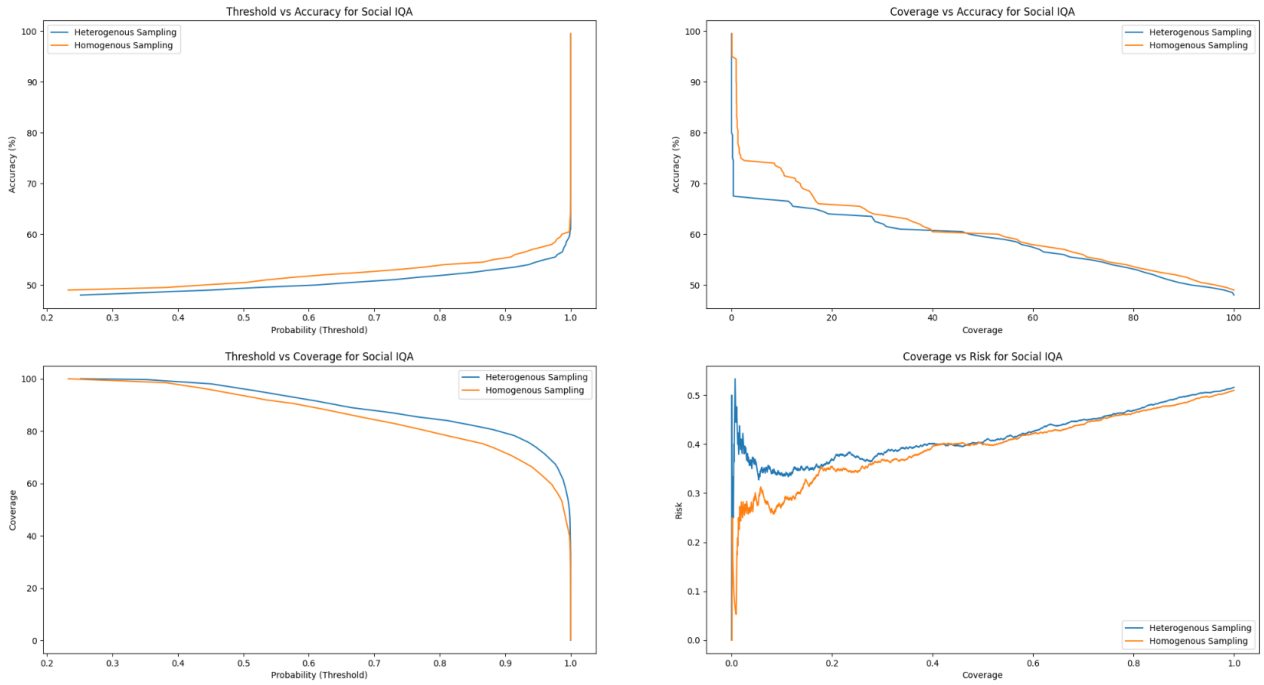

Fig 7: Selective Prediction Metrics Plots for Social IQA (Homogenous and Heterogenous Sampling)

Literature Review

We went through the papers 1-4 to understand selective prediction itself and how models can be calibrated to abstain from answering if they are not confident enough. For understanding Multi-Task Learning, how to combine different types of datasets and the dataset sampling strategies for training a multi-Task model, we read the papers and articles 5-7.

- [1] Selective Question Answering under Domain Shift - https://arxiv.org/abs/2006.09462
- [2] It's better to say "I can't answer" than answering incorrectly: Towards Safety critical NLP systems - https://arxiv.org/abs/2008.09371
- [3] The Art of Abstention: Selective Prediction and Error Regularization for Natural Language Processing - https://aclanthology.org/2021.acl-long.84/
- [4] Will this Question be Answered? Question Filtering via Answer Model Distillation for Efficient Question Answering - https://arxiv.org/abs/2109.07009
- [5] Dynamic Sampling Strategies for Multi-Task Reading Comprehension - https://aclanthology.org/2020.acl-main.86/
- [6] The Natural Language Decathlon: Multitask Learning as Question Answering - https://arxiv.org/abs/1806.08730
- [7] A Primer on Multi-Task Learning https://medium.com/analytics-vidhya/a-primer-on-multi-task-learning-in-nlp-part-1-7154b4227c0e