Please submit a report describing your final project. A good project report should include the following:

**· Task Definition and Motivation**

We tried to solve the SQUAD 2.0 dataset, which is a question-answering problem with unanswerable questions.

When we start the project, we realized many existing methods had achieved excellent precision scores on the leaderboard. However, we noticed that most of the models are BERT like models, which are extremely large. In this project, we tried to find some reasonable models and further downscale it and keep it have similar accuracy.

**· Related Work**

In our test, we used fine-tuning method instead of retraining the entire model. From our search, we notice that the XLNet and Albert are the two reasonable models use. Other Bert like models have too many parameters for our GPUs and takes too much training time.

Both XLNet and Albert are Bert like models. XLNet used single direction information during the encoder phase and proposed the permutation language model by combing Auto Regression and Auto Encoder. Alber is an improvement of Bert model, which factorized embedding parameters, used cross-layer parameter sharing, and used sentence order prediction instead of next sentence prediction.

* Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. arXiv preprint arXiv:1909.11942, 2019.

* Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. XLNet: Generalized autoregressive pretraining for language understanding. arXiv preprint arXiv:1906.08237, 2019.

**· Proposed Approach**

In our test, we used fine-tuning method instead of retraining the entire model. After deciding to use XLNet and Albert as our initial model. We used proposed using different hyperparameters to downscale the model and train the model for fine-tuning purposes. Most of our tests did not work out will show some useful results in the next sections.

**· Dataset and Evaluation Measures description**

We used SQUAD 2.0 as our dataset. Since the dataset contains answerable and unanswerable questions, we measure the model based on the overall F1 score, question has answer F1 score, and the question has no answer F1 score.

**· Experimental Results and Analysis of Results**

Like we mentioned before, most of the experiments did not work. The biggest failure is the entire downscale process of the XLNet model. One of our key idea of the downscale process is by reducing the max sequence length of the input, because it can dramatically reduce the model parameters and training time. However, XLNet is extremely sensitive to such change, and the performance reduce from F1 score 81.2 to 38.34, when changing max sequence length from 384 to 256. We tried to recover such an issue by longer training time, but the model still has terrible performance. At that point, we have to claim that we are not able to downscale XLNet.

Albert, on the other hand, works very well, as the image shows. When we reduce the max length from 348 to 100/128, the overall F1 score is still higher than 70. Notice that the model parameters have been reduced for about 60%, and the training time per epoch has been reduced from two hours to forty minutes.

We believe that changing the model in this way has limitations. Because the model has been majorly downscaled, the capability of the model is also reduced. The model has a hard time to getting better results by having more training epochs or using larger pre-trained models.

**· A section about what everyone on the team worked on**

In our team, we have three members. Minghao Liu (First-year master student), Cagan Bakirci (Fourth-year Undergraduate student), and Jing Xiong (First year PhD student). At the beginning of the quarter, Jing Xiong decided to drop the class, so there are officially two members in our team. Jing Xion still chose to support our project after she dropped the class. Thanks to Jing Xiong, we get a GTX 1080 GPU for our project.

In the project, start from finding models, setting things up, training models, to getting different results. Minghao and Cagan worked on all parts of them. Since Minghao has more prior knowledge in Machine Learning, Minghao is more productive than Cagan. However, there is no doubt both team members tried their best. Hence, we would like to call it even work.