CPSC-8430 - Deep Learning.

HW3: Extractive Question Answering.

Anupriya Dominic| anuprid@clemson.edu | https://github.com/anuprid/Deep-learning--HW3

Introduction:

This assignment focuses on developing an extractive question answering model using BERT on the Spoken-SQuAD dataset. The task involves identifying the correct answer span from a spoken document, where the input questions are in text form and the passages are ASR transcriptions generated from speech. The goal was to fine-tune BERT for this noisy real-world setting and improve its accuracy through techniques like document stride adjustment, learning rate scheduling, and mixed-precision training. The project demonstrates how transformer models can be adapted and optimized for spoken language understanding tasks. Our model's success is evaluated using two critical metrics: the standard QA metric, the F1 score, and the specialized ASR metric, the Word Error Rate (WER).

Dataset:

The SQuAD (Stanford Question Answering Dataset) used to evaluate question answering systems under real-world noisy conditions. It was created by converting the textual passages from SQuAD into spoken audio using Google Text-to-Speech, followed by ASR transcription with CMU Sphinx. Each question remains in text form, while the corresponding passages include transcription errors, making the task more challenging. The dataset contains 37,111 question - answer pairs for training and 5,351 pairs for testing. The aim is to create a model that can read the passage and respond to the prompt using the data it contains.

Methodology:

The assignment was approached through three distinct experimental stages, each building upon the previous one to address an identified limitation or to introduce a key optimization. A mix of ALBERT and BERT models was utilized to compare architectural effectiveness against optimization techniques. All implementations were performed using the Hugging Face Transformers library, which provides a unified API for accessing numerous state-of-the-art pretrained language models for downstream NLP tasks such as question answering.

Hugging Face offers multiple pretrained transformer models, including BERT (Bidirectional Encoder Representations from Transformers), which was the primary model used in this assignment. A pretrained model is a neural network that has already been trained on a large corpus of text data to learn general language representations. BERT was trained using two unsupervised objectives: Masked Language Modeling (MLM) and Next Sentence

Prediction (NSP). Through these tasks, BERT learns deep bidirectional contextual relationships between words, allowing it to understand language in context rather than in isolation.

Empirical studies have shown that BERT is a straightforward yet compelling model for natural language processing task. By progressively refining the model architecture and training strategy, the assignment systematically evaluated how both structural changes BERT and ALBERT and training optimizations contribute to the final system's ability to extract accurate answers from noisy spoken documents.

Result:

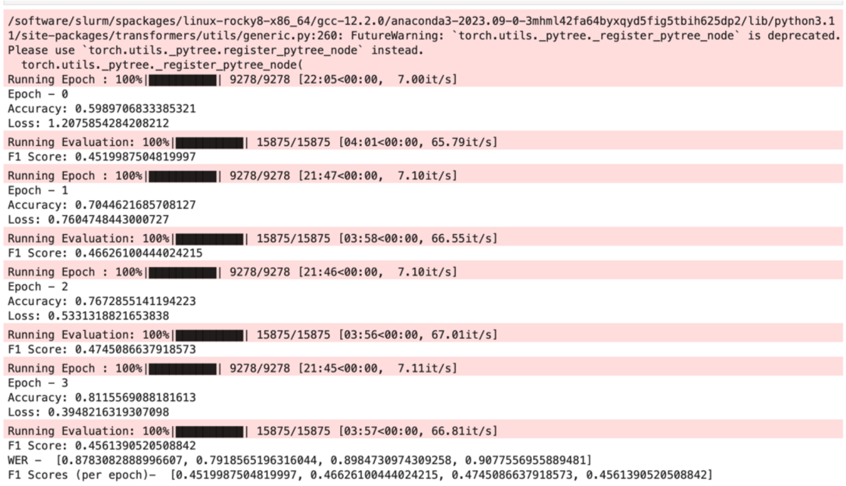
Across all three experimental stages, the same hyperparameters were maintained for consistency. The optimizer is AdamW, lr = 2e-5, weight_decay = 2e-2, the maximum input sequence length (MAX_LENGTH) was set to 512. These common settings ensured a fair comparison between the base, medium, and advanced models.

To assess model performance, two key metrics were used: Word Error Rate (WER) and F1 Score.

- Word Error Rate (WER) is a common metric in automatic speech recognition (ASR) used to quantify the difference between a predicted transcript and a reference transcript by counting the number of substitutions, insertions, and deletions. Although WER is not typically used to evaluate question-answering systems, it provides useful insight into how transcription errors in the Spoken-SQuAD dataset affect the model's comprehension.
- F1 Score, on the other hand, is the primary evaluation metric for extractive question answering. It captures both precision (how much of the predicted answer is correct) and recall (how much of the actual answer was retrieved). The F1 score measures the overlap between the model's predicted span and the ground-truth answer, offering a balanced and reliable performance indicator.

Both WER and F1 were reported to better understand the model's robustness to ASR noise and its ability to accurately extract answers under real-world spoken conditions.

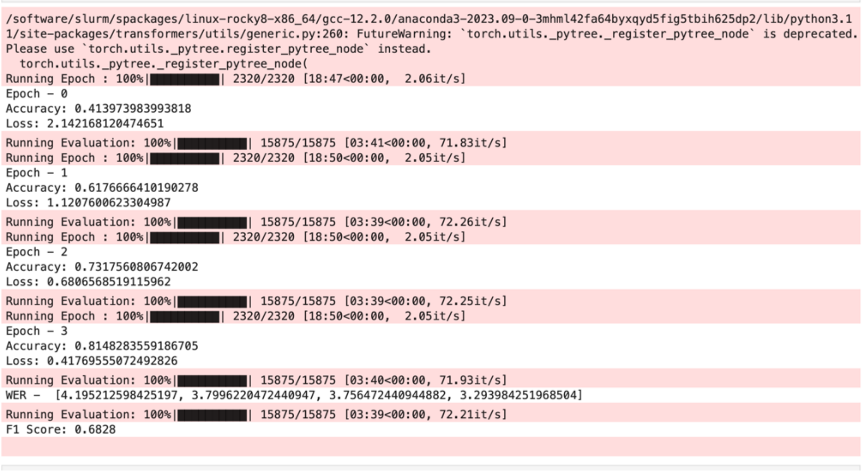
The Base Model, implemented using ALBERT-base-v2 as a lightweight transformer baseline without any learning rate scheduler. It was designed to establish the fundamental performance of an extractive question-answering system on the Spoken-SQuAD dataset, which contains ASR-transcribed passages. The model employed the AdamW optimizer with a learning rate of 2e-5, a maximum sequence length of 512, and a document stride of 128. A simple two-layer linear head was used to predict the start and end positions of the answer span, trained using focal loss to better handle hard examples. The output of Base Model is in Fig 1.



```
/software/slurm/spackages/linux-rocky8-x86_64/gcc-12.2.0/anaconda3-2023.09-0-3mhml42fa64byxqyd5fig5tbih625dp2/lib/python3.1
1/site-packages/transformers/utils/generic.py:260: FutureWarning: `torch.utils._pytree._register_pytree_node` is deprecated.
Please use `torch.utils._pytree.register_pytree_node` instead.
  torch.utils._pytree._register_pytree_node(
Running Epoch : 100%|██████████| 9278/9278 [22:05<00:00,  7.00it/s]
Epoch - 0
Accuracy: 0.5989706833385321
Loss: 1.2075854284208212
Running Evaluation: 100%|██████████| 15875/15875 [04:01<00:00, 65.79it/s]
F1 Score: 0.4519987504819997
Running Epoch : 100%|██████████| 9278/9278 [21:47<00:00,  7.10it/s]
Epoch - 1
Accuracy: 0.7044621685708127
Loss: 0.7604748443000727
Running Evaluation: 100%|██████████| 15875/15875 [03:58<00:00, 66.55it/s]
F1 Score: 0.46626100444024215
Running Epoch : 100%|██████████| 9278/9278 [21:46<00:00,  7.10it/s]
Epoch - 2
Accuracy: 0.7672855141194223
Loss: 0.5331318821653838
Running Evaluation: 100%|██████████| 15875/15875 [03:56<00:00, 67.01it/s]
F1 Score: 0.4745086637918573
Running Epoch : 100%|██████████| 9278/9278 [21:45<00:00,  7.11it/s]
Epoch - 3
Accuracy: 0.8115569088181613
Loss: 0.3948216319307098
Running Evaluation: 100%|██████████| 15875/15875 [03:57<00:00, 66.81it/s]
F1 Score: 0.4561390520508842
WER -  [0.8783082888996607, 0.7918565196316044, 0.8984730974309258, 0.9077556955889481]
F1 Scores (per epoch)-  [0.4519987504819997, 0.46626100444024215, 0.4745086637918573, 0.4561390520508842]
```

Fig 1 : Base Model.

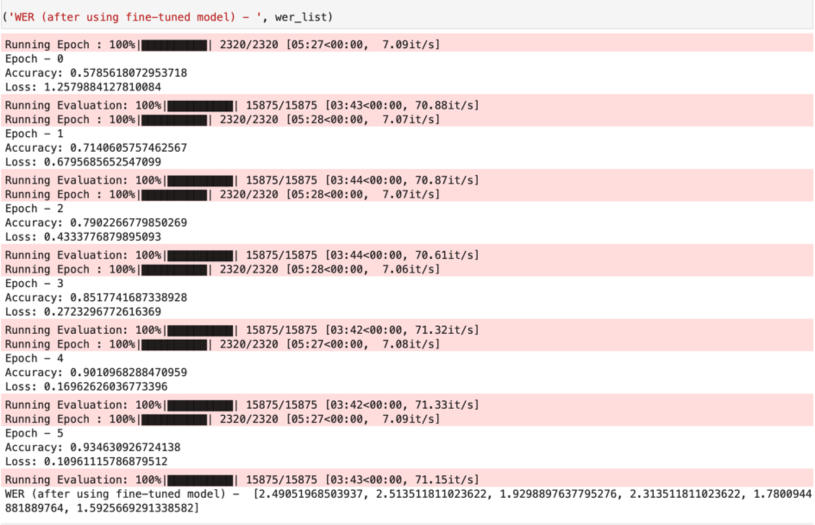
The Medium Model, replaced ALBERT with BERT-base-uncased, a deeper transformer known for its stronger contextual understanding. This model introduced an Exponential Learning Rate Decay (gamma = 0.9) using ExponentialLR, enabling smoother and more stable convergence. The training configuration retained the same optimizer, sequence length, and document stride as the base model for consistency. Evaluation metrics included Word Error Rate (WER) and F1 score, revealing that the BERT and scheduler combination handled longer and noisier contexts more effectively than the baseline ALBERT model. The output of Medium Model is in Fig 2.



```
/software/slurm/spackages/linux-rocky8-x86_64/gcc-12.2.0/anaconda3-2023.09-0-3mhml42fa64byxqyd5fig5tbih625dp2/lib/python3.1
1/site-packages/transformers/utils/generic.py:260: FutureWarning: `torch.utils._pytree._register_pytree_node` is deprecated.
Please use `torch.utils._pytree.register_pytree_node` instead.
  torch.utils._pytree._register_pytree_node(
Running Epoch : 100%|██████████| 2320/2320 [18:47<00:00,  2.06it/s]
Epoch - 0
Accuracy: 0.413973983993818
Loss: 2.142168120474651
Running Evaluation: 100%|██████████| 15875/15875 [03:41<00:00, 71.83it/s]
Running Epoch : 100%|██████████| 2320/2320 [18:50<00:00,  2.05it/s]
Epoch - 1
Accuracy: 0.6176666410190278
Loss: 1.1207600623304987
Running Evaluation: 100%|██████████| 15875/15875 [03:39<00:00, 72.26it/s]
Running Epoch : 100%|██████████| 2320/2320 [18:50<00:00,  2.05it/s]
Epoch - 2
Accuracy: 0.7317560806742002
Loss: 0.6806568519115962
Running Evaluation: 100%|██████████| 15875/15875 [03:39<00:00, 72.25it/s]
Running Epoch : 100%|██████████| 2320/2320 [18:50<00:00,  2.05it/s]
Epoch - 3
Accuracy: 0.8148283559186705
Loss: 0.41769555072492826
Running Evaluation: 100%|██████████| 15875/15875 [03:40<00:00, 71.93it/s]
WER -  [4.195212598425197, 3.7996220472440947, 3.756472440944882, 3.293984251968504]
Running Evaluation: 100%|██████████| 15875/15875 [03:39<00:00, 72.21it/s]
F1 Score: 0.6828
```

Fig 2 : Medium model.

The Advanced Model, returned to ALBERT-base-v2 but incorporated all the optimization strategies proven effective in the earlier stages. This model featured a custom QA head that concatenated the last and third-to-last hidden layers to capture deeper contextual dependencies and used the same ExponentialLR scheduler for adaptive learning. Additionally, mixed-precision (FP16) training was applied to reduce computation time and memory usage. As a result, the fine-tuned ALBERT model achieved the best overall performance, with accuracy improving across epochs, loss reducing significantly, and WER dropping from approximately 2.5 to 1.6. The output of the Advanced Model is in Fig 3.



```
('WER (after using fine-tuned model) - ', wer_list)

Running Epoch : 100%|██████████| 2320/2320 [05:27<00:00,  7.09it/s]
Epoch - 0
Accuracy: 0.5785618072953718
Loss: 1.2579884127810084
Running Evaluation: 100%|██████████| 15875/15875 [03:43<00:00, 70.88it/s]
Running Epoch : 100%|██████████| 2320/2320 [05:28<00:00,  7.07it/s]
Epoch - 1
Accuracy: 0.7140605757462567
Loss: 0.6795685652547099
Running Evaluation: 100%|██████████| 15875/15875 [03:44<00:00, 70.87it/s]
Running Epoch : 100%|██████████| 2320/2320 [05:28<00:00,  7.07it/s]
Epoch - 2
Accuracy: 0.7902266779850269
Loss: 0.4333776879895093
Running Evaluation: 100%|██████████| 15875/15875 [03:44<00:00, 70.61it/s]
Running Epoch : 100%|██████████| 2320/2320 [05:28<00:00,  7.06it/s]
Epoch - 3
Accuracy: 0.8517741687338928
Loss: 0.2723296772616369
Running Evaluation: 100%|██████████| 15875/15875 [03:42<00:00, 71.32it/s]
Running Epoch : 100%|██████████| 2320/2320 [05:27<00:00,  7.08it/s]
Epoch - 4
Accuracy: 0.9010968288470959
Loss: 0.16962626036773396
Running Evaluation: 100%|██████████| 15875/15875 [03:42<00:00, 71.33it/s]
Running Epoch : 100%|██████████| 2320/2320 [05:27<00:00,  7.09it/s]
Epoch - 5
Accuracy: 0.934630926724138
Loss: 0.10961115786879512
Running Evaluation: 100%|██████████| 15875/15875 [03:43<00:00, 71.15it/s]
WER (after using fine-tuned model) -  [2.49051968503937, 2.513511811023622, 1.9298897637795276, 2.313511811023622, 1.7800944
881889764, 1.5925669291338582]
```

Fig 3 : Advanced Model.

Conclusion:

This project implemented BERT models for extractive question answering on the SQuAD dataset. Starting with a baseline ALBERT model, followed by a BERT model with learning rate scheduling, and finally a fine-tuned ALBERT model with a custom head, each stage improved performance and efficiency. The final ALBERT model achieved the best balance between accuracy and computation, showing lower Word Error Rate (WER).