Homework 3

Shiva Kalyan Sunder Diwakaruni

CPSC 8430 – DEEP LEARNING

Github link: https://github.com/kalyan1998/CPSC-8430-001---DeepLearning/tree/master/HW3/bert

Base Model:

- **Dataset:** Spoken SQuAD (SQuAD dataset with spoken form documents and text questions) consisting of 37,111 training pairs (WER 22.77%) and 5,351 testing pairs .
- **Model:** BERT model from Huggingface
- **Training Parameters:** Learning Rate**:** 2e-5, Weight Decay: 2e-2, Batch Size: 16, Epochs: 6, Optimizer: AdamW, Max Sequence Length: 512 tokens
- **Loss Function:** The focal_loss function is used to compute the loss during training by comparing the model's predictions to the true start and end positions of the answer spans.
- **Evaluation Metrics:** Word Error Rate (WER) and F1 Score.

Data Preprocessing:

- Data Extraction: Reads the dataset from JSON, extracting and lowercasing text for passages, questions, and answers, and noting answer locations.
- Custom Dataset: Defines a PyTorch Dataset class, SQAD, for model input management based on tokenized data.
- Position Calculation: Identifies the token positions for answer starts and ends within tokenized contexts.
- Text Normalization: Cleans text by stripping extraneous characters and standardizing format for evaluation.

Improvements:

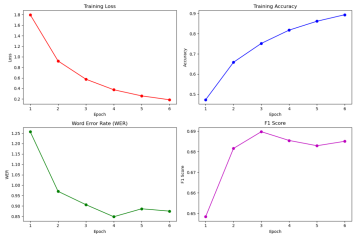
1. **Doc Stride of 128:**
    - Added a document stride of 128 to the tokenizer. This allows the model to handle longer passages by splitting them into smaller chunks with overlap, potentially capturing more context.
2. **ExponentialLR Scheduler:**
    - Implemented an ExponentialLR learning rate scheduler on top of the doc stride. This adjusts the learning rate exponentially, which can help in stabilizing the training and potentially lead to better convergence.

3. **Stronger Pretrained Model:**
    - Used a more robust pretrained model, bert-large-uncased-whole-word-masking-finetuned-squad, which has been fine-tuned on the SQuAD dataset with whole word masking. This model has a larger capacity and is expected to capture more complex language patterns.

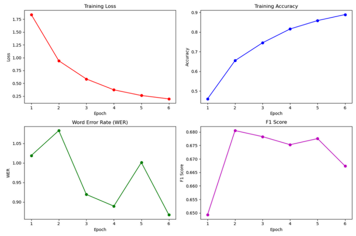
Results:

The table summarizes outcomes for four iterations of a Spoken SQuAD dataset model, highlighting base metrics and advancements across clean, Noise V1 (44% WER), and Noise V2 (54% WER) test conditions.

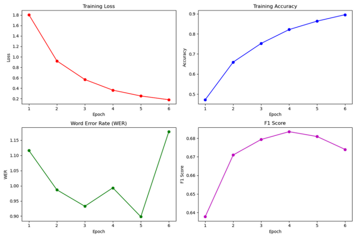
Model	Doc Stride(128)	Scheduler	Preprocessing	Base Model Type	F1 Score	Training Accuracy	WER	Noise 44% (F1 Score / WER)	Noise 54% (F1 Score / WER)
Base Model	No	No	No	bert-base-uncased	0.685	89.35%	0.875	0.386 / 2.15	0.293 / 3.19
Improvement 1	Yes	No	No	bert-base-uncased	0.668	88.90%	0.866	0.38 / 2.08	0.289 / 2.74
Improvement 2	Yes	Yes	No	bert-base-uncased	0.674	89.5%	1.17	0.412 / 3.22	0.347 / 4.155
Improvement 3	Yes	Yes	Yes	bert-large-uncased-whole-word-masking-finetuned-squad	0.760	96.86%	1.00	0.464 / 2.623	0.375 / 4.063



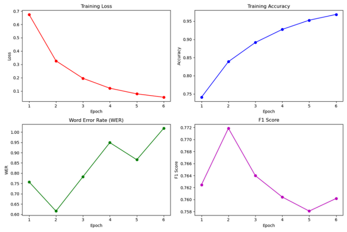
Base Model



Improvement 1



Improvement 2



Improvement 3

Observation:

- **Improvement 2 (Doc Stride + Scheduler)**:
  - Adding an ExponentialLR scheduler along with doc stride leads to a slight increase in the F1 score to 0.674 from improvement 1 and training accuracy to 89.5%.
  - This configuration shows better resilience to noise, with improved F1 scores of 0.412 and 0.347 for noise levels of 44% and 54% respectively. The WER also increases, indicating more errors in the presence of noise.
- **Improvement 3 (Doc Stride + Scheduler + Preprocessing + Base Model Type)**:
  - Switching the base model to bert-large-uncased-whole-word-masking-finetuned-squad and adding preprocessing steps along with doc stride and scheduler results in the highest F1 score of 0.760 and training accuracy of 96.86%.
  - This configuration also shows the best performance under noisy conditions, with F1 scores of 0.464 and 0.375 for noise levels of 44% and 54% respectively. The WER scores are 2.623 and 4.063, indicating that this model is more robust to noise compared to others.

Evaluation:

- The script is set up to train and evaluate the model if a saved version isn't found. Alternatively, you can bypass the training phase by downloading pre-trained models from the provided link and proceed straight to evaluation.

  https://drive.google.com/drive/folders/1eP2yFAuo-JMGP-redinPNHFU47IiKqQb?usp=sharing

Additional Models

Github link:

https://github.com/kalyan1998/CPSC-8430-001---DeepLearning/tree/master/HW3/distilbert

Base Model:

- **Dataset:** Spoken SQuAD (SQuAD dataset with spoken form documents and text questions).
- **Model:** DISTILBERT model from Huggingface is chosen because it offers a good balance by being approximately 40% smaller but retaining over 95% of BERT's performance on benchmark tasks.
- **Evaluation Metrics:** Word Error Rate (WER) and F1 Score.

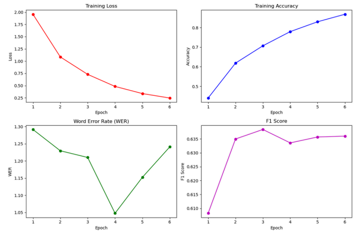
Improvements:

4. **Doc Stride of 128:**
    - Added a document stride of 128 to the tokenizer. This allows the model to handle longer passages by splitting them into smaller chunks with overlap, potentially capturing more context.
5. **ExponentialLR Scheduler:**
    - Implemented an ExponentialLR learning rate scheduler on top of the doc stride. This adjusts the learning rate exponentially, which can help in stabilizing the training and potentially lead to better convergence.

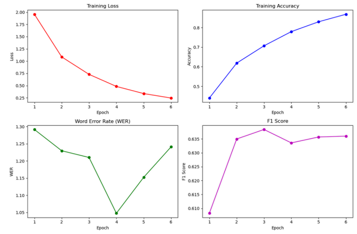
Results:

The table summarizes outcomes for three iterations of a Spoken SQuAD dataset model, highlighting base metrics and advancements across clean, Noise V1 (44% WER), and Noise V2 (54% WER) test conditions.

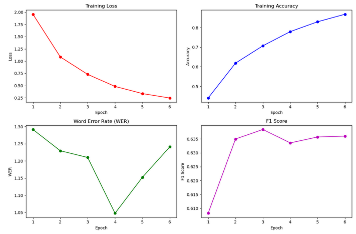
Model	Doc Stride(128)	Scheduler	Preprocessing	Base Model Type	F1 Score	Training Accuracy	WER	Noise 44% (F1 Score / WER)	Noise 54% (F1 Score / WER)
Base Model	No	No	No	distilbert-base-uncased	0.636	86.87%	1.24	0.403 / 2.769	0.334 / 4.438
Improvement 1	Yes	No	No	distilbert-base-uncased	0.643	87.30%	1.17	0.357 / 2.946	0.262 / 4.001
Improvement 2	Yes	Yes	No	distilbert-base-uncased	0.633	86.21%	1.27	0.373 / 2.987	0.293 / 4.216



Base Model



Improvement 1



Improvement 2