Assignment No.02

Deep Learning

Name: Shahid Ullah

Roll No.: 22I-2019

Section: DS-A

Legal Clause Similarity

1. Network Details and Baseline Rationale (10 pts)

Dataset Overview

- **Total Clauses:** 150,881
- **Unique Categories:** 395
- **Top Categories:** *time-of-essence*, *definitions-and-interpretation*, *capitalized-terms*, *headings*, *exhibits*, *definitions*, etc.
- **Vocabulary Size:** 44,812
- **Training Pairs:** 16,000
- **Validation Pairs:** 4,000
- **Batch Size:** 64

Model Architectures

Two deep learning baselines were implemented for semantic similarity classification between legal clauses:

1. **BiLSTM Model**
    - **Embedding Layer:** Pretrained word embeddings (trainable)
    - **Encoder:** Bidirectional LSTM capturing contextual dependencies
    - **Pooling:** Max and average pooling for sentence-level representation
    - **Classifier:** Fully connected layers with ReLU and sigmoid activation
    - **Total Parameters:** 14,344,977
2. **Attention-Based BiLSTM**
    - Same structure as BiLSTM with an **attention layer** to emphasize contextually important words.
    - **Total Parameters:** 14,345,234

Training Configuration

- **Optimizer:** Adam
- **Loss Function:** Binary Cross-Entropy
- **Mixed Precision Training:** Enabled using torch.cuda.amp

- **Early Stopping:** Applied based on validation loss
- **Epochs:** 10
- **Device:** CUDA GPU

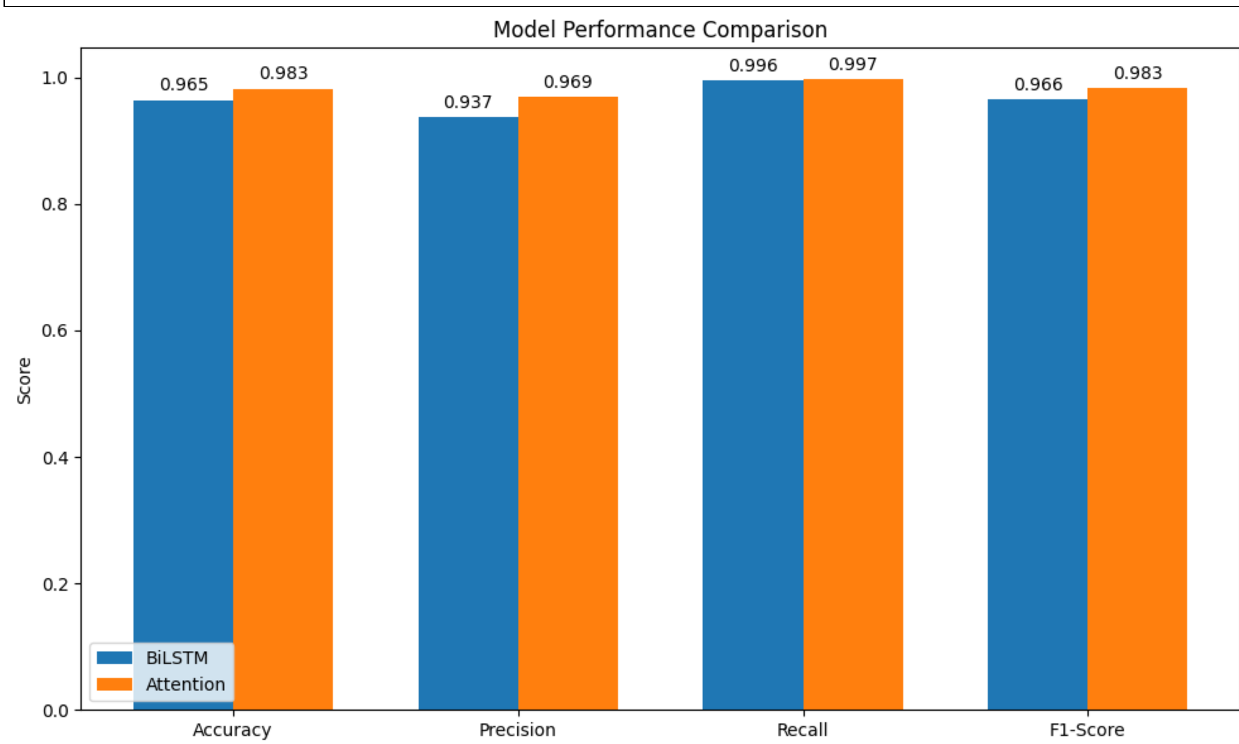
Rationale for Choosing Baselines

Legal text similarity requires capturing *semantic context* rather than surface word matching.

- **BiLSTM** was chosen as a strong sequential baseline for capturing bidirectional dependencies in legal clauses.
- **Attention-BiLSTM** was introduced as an enhancement to allow the model to weigh key legal terms more heavily — a critical factor in clause interpretation.

2. Comparison Between Baselines (10 pts)

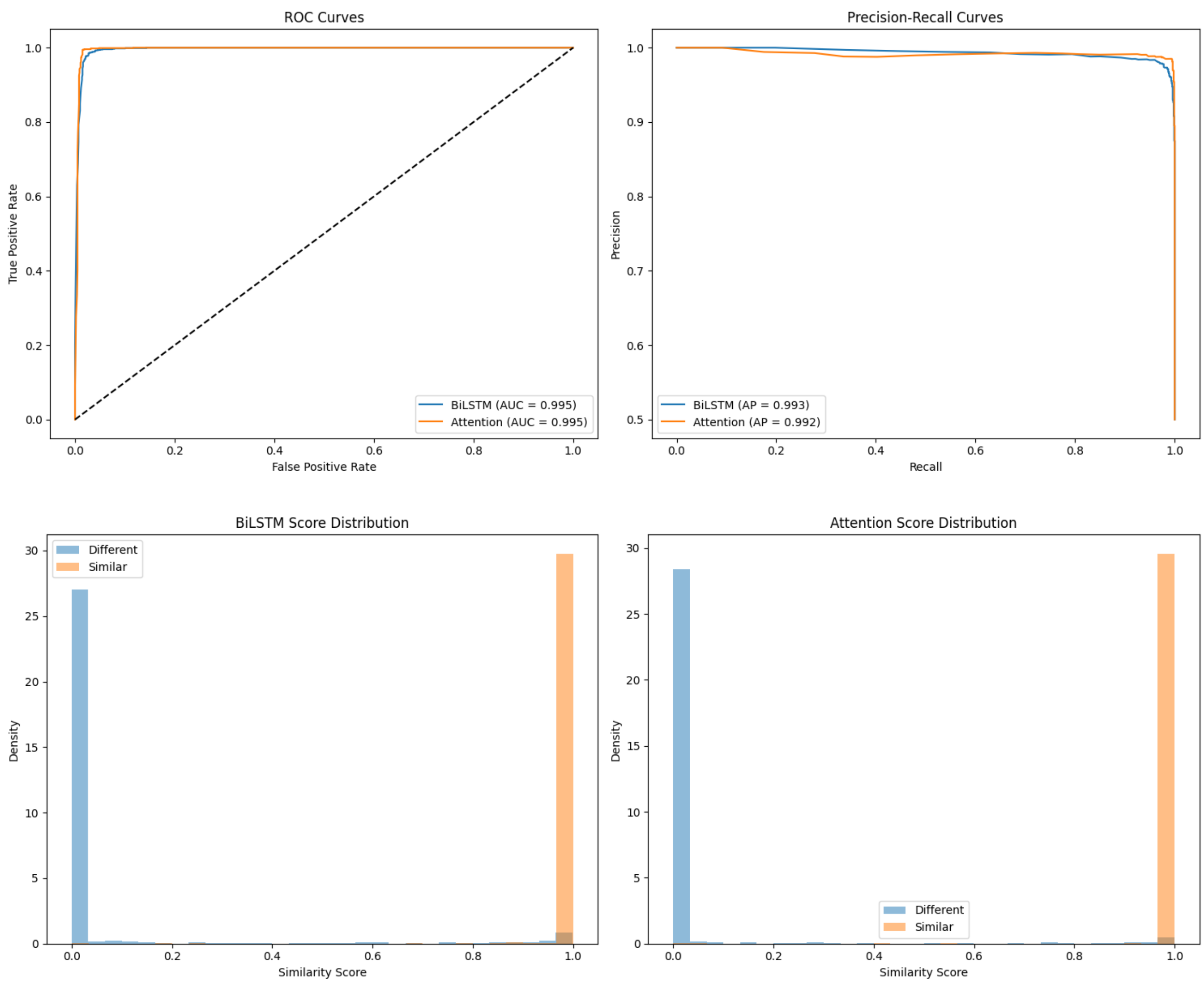
Model	Train Accuracy	Validation Accuracy	Train Loss	Validation Loss	Test Accuracy	Precision	Recall	F1-Score	ROC-AUC	PR-AUC
BiLSTM	0.9916	0.9643	0.0294	0.1384	0.9645	0.9370	0.996	0.9656	0.9949	0.9929
Attention -BiLSTM	0.9942	0.9826	0.0197	0.0582	0.9825	0.9689	0.997	0.9828	0.9952	0.9918

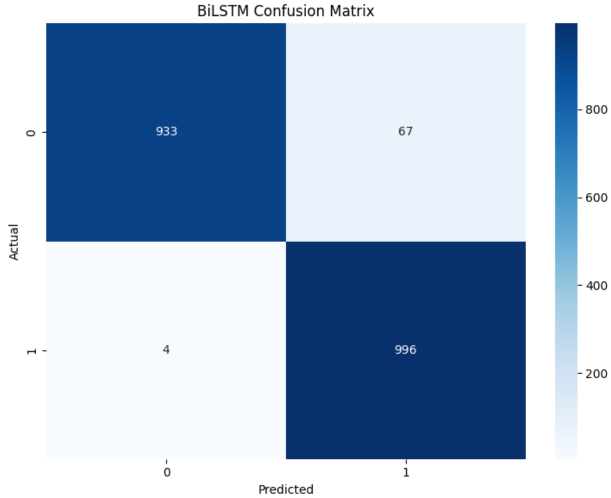


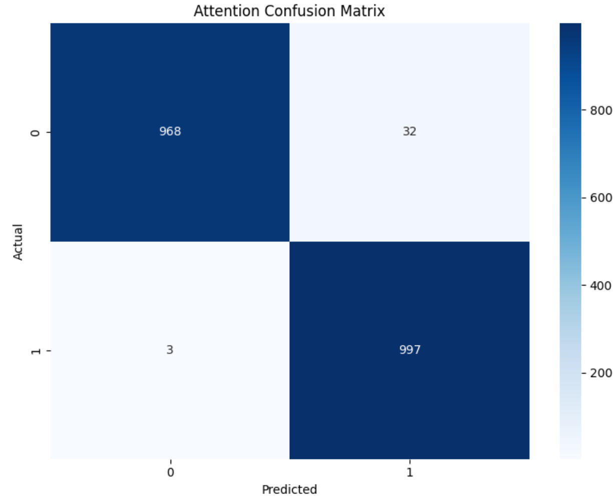
Comparison Summary:

- Both models achieve excellent performance, but the **Attention model** consistently outperforms BiLSTM by a margin of ~1.8% in validation accuracy and 1.7% in F1-score.
- The attention mechanism improves interpretability and generalization, particularly useful for clauses with long, complex syntax.

3. Graphs (10 pts)







4. Performance Measures and Domain Evaluation Discussion (15 pts)

Evaluation Metrics Used

- **Accuracy:** Overall percentage of correctly predicted clause pairs.
- **Precision:** Measures how many predicted similar pairs were truly similar — crucial for avoiding false positives in contract similarity detection.
- **Recall:** Indicates how well the model identifies all truly similar clauses — important in ensuring no critical legal similarity is missed.
- **F1-Score:** Harmonic mean of precision and recall; ideal when both false positives and negatives are costly.
- **ROC-AUC:** Captures how well the model separates similar vs dissimilar clauses across thresholds.
- **PR-AUC:** Important when dataset imbalance exists; measures performance in positive (similar) class detection.

Metric Interpretation

- Both models have high recall (>0.99), showing they correctly identify almost all similar clauses.
- Precision (~0.94 for BiLSTM and ~0.97 for Attention) ensures few incorrect matches.
- High F1 and ROC-AUC (>0.98) confirm strong generalization.

Most Suitable Metric "In the Wild"

In real-world legal systems, **high precision** is the most crucial metric.
A system suggesting clause similarities should avoid false matches, as incorrect legal equivalence can cause serious interpretation issues.
Thus, while recall and F1 are strong, **Precision and ROC-AUC** are most critical for deployment in the wild.

5. Examples of Correctly and Incorrectly Matched Clauses (4 pts)

Correctly Matched Clauses (Similar Meaning):

1. *"This agreement shall be governed by the laws of the State of New York."*
   "The contract shall be interpreted under New York State law."
2. *"Each party agrees to maintain confidentiality of all shared data."*
   "Both parties must ensure confidentiality of exchanged information."

Incorrectly Matched Clauses (False Positives):

1. *"The supplier shall deliver goods within thirty days."*
   "Either party may terminate this agreement upon breach."
2. *"All disputes shall be settled through arbitration."*
   "This clause defines the governing jurisdiction of the contract."

These illustrate how the model sometimes confuses legal procedural terms (e.g., *arbitration* vs *jurisdiction*).

Conclusion:

Both **BiLSTM** and **Attention-based BiLSTM** architectures performed strongly for legal clause similarity detection, with the attention variant showing superior accuracy and interpretability. Given the high precision and ROC-AUC values, the Attention model is best suited for a real-world legal document comparison system, capable of distinguishing nuanced contractual meanings with reliability.